Structure Learning of Gene Interaction Networks

by

Piet Jones

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in the Science at Stellenbosch University

Department of Mathematics,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisors:

Dr. D. Jacobson Dr. P. Grobler

January 2014

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: .....

Abstract

Structure Learning of Gene Interaction Networks

P. Jones

Department of Mathematics,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc (Mathematics)

January 2014

There is an ever increasing wealth of information that is being generated regarding biological systems, in particular information on the interactions and dependencies of genes and their regulatory process. It is thus important to be able to attach functional understanding to this wealth of information. Mathematics can potentially provide the tools needed to generate the necessary abstractions to model the complex system of gene interaction.

Here the problem of uncovering gene interactions is cast in several contexts, namely uncovering gene interaction patterns using statistical dependence, co-occurrence as well as feature enrichment. Several techniques have been proposed in the past to solve these, with various levels of success. Techniques have ranged from supervised learning, clustering analysis, boolean networks to dynamical Bayesian models and complex system of differential equations. These models attempt to navigate a high dimensional space with challenging degrees of freedom.

In this work a number of approaches are applied to hypothesize a gene interaction network structure. Three different models are applied to real biological data to generate hypotheses on putative biological interactions. A cluster-based analysis combined with a feature enrichment detection is initially applied to a *Vitis vinifera* dataset, in a targetted analysis. This model bridges a disjointed set of putatively co-expressed genes based on significantly associated features, or experimental conditions. We then apply a cross-cluster Markov Blanket based model, on a *Saccharomyces cerevisiae* dataset. Here the disjointed clusters are bridged by estimating statistical dependence relationship across clusters, in an un-targetted approach. The final model applied to the same *Saccharomyces cerevisiae* dataset is a non-parametric Bayesian

method that detects probeset co-occurrence given a local background and inferring gene interaction based on the topological network structure resulting from gene co-occurance. In each case we gather evidence to support the biological relevance of these hypothesized interactions by investigating their relation to currently established biological knowledge.

The various methods applied here appear to capture different aspects of gene interaction, in the datasets we applied them to. The targetted approach appears to putatively infer gene interactions based on functional similarities. The cross-cluster-analysis-based methods, appear to capture interactions within pathways. The probabilistic-co-occurrence-based method appears to generate modules of functionally related genes that are connected to potentially explain the underlying experimental dynamics.

Uittreksel

Struktuur Leer van Interaksie Netwerke van Gene

("Structure Learning of Gene Interaction Networks")

P. Jones

Departement Wiskunde,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc (Wiskunde)

Januarie 2014

Daar is 'n toenemende rykdom van inligting wat gegenereer word met betrekking tot biologiese stelsels, veral inligting oor die interaksies en afhanklikheidsverhoudinge van gene asook hul regulatoriese prosesse. Dit is dus belangrik om in staat te wees om funksionele begrip te kan heg aan hierdie rykdom van inligting. Wiskunde kan moontlik die gereedskap verskaf en die nodige abstraksies bied om die komplekse sisteem van gene interaksies te modelleer.

Hier is die probleem met die beraming van die interaksies tussen gene benader uit verskeie kontekste uit, soos die ontdekking van patrone in gene interaksie met behulp van statistiese afhanklikheid , mede-voorkoms asook funksie verryking. Verskeie tegnieke is in die verlede voorgestel om hierdie probleem te benader, met verskillende vlakke van sukses. Tegnieke het gewissel van toesig leer , die groepering analise, boolean netwerke, dinamiese Bayesian modelle en 'n komplekse stelsel van differensiaalvergelykings. Hierdie modelle poog om 'n hoë dimensionele ruimte te navigeer met uitdagende grade van vryheid.

In hierdie werk word 'n aantal benaderings toegepas om 'n genetiese interaksie netwerk struktuur voor te stel. Drie verskillende modelle word toegepas op werklike biologiese data met die doel om hipoteses oor vermeende biologiese interaksies te genereer. 'n Geteikende groeperings gebaseerde analise gekombineer met die opsporing van verrykte kenmerke is aanvanklik toegepas op 'n *Vitis vinifera* datastel. Hierdie model verbind disjunkte groepe van vermeende mede-uitgedrukte gene wat gebaseer is op beduidende verrykte kenmerke, hier eksperimentele toestande . Ons pas dan 'n tussen groepering Markov Kombers model toe, op 'n *Saccharomyces cerevisiae* datastel. Hier is die disjunkte groeperings ge-oorbrug deur die beraming van statistiese afhanklikheid verhoudings

tussen die elemente in die afsondelike groeperings. Die finale model was ons toepas op dieselfde *Saccharomyces cerevisiae* datastel is 'n nie- parametriese Bayes metode wat probe stelle van mede-voorkommende gene ontdek, gegee 'n plaaslike agtergrond. Die gene interaksie is beraam op grond van die topologie van die netwerk struktuur veroorsaak deur die gesamentlike voorkoms gene. In elk van die voorgenome gevalle word ons hipotese vermoedelik ondersteun deur die beraamde gene interaksies in terme van huidige biologiese kennis na te vors.

Die verskillende metodes wat hier toegepas is, modelleer verskillende aspekte van die interaksies tussen gene met betrekking tot die datastelle wat ons ondersoek het. In die geteikende benadering blyk dit asof ons vermeemde interaksies beraam gebaseer op die ooreenkoms van biologiese funksies. Waar die afleide gene interaksies moontlik gebaseer kan wees op funksionele ooreenkomste tussen die verskeie gene. In die analise gebaseer op die tussen modelering van gene groepe, blyk dit asof die verhouding van gene in bekende biologiese substelsels gemodelleer word. Dit blyk of die model gebaseer op die gesamentlike voorkoms van gene die verband tussen groepe van funksionele verbonde gene modelleer om die onderliggende dinamiese eienskappe van die experiment te verduidelik.

Acknowledgements

I would like to acknowledge the support of my supervisor, Dan Jacobson, I am thankful that he still tolerates my idiosyncrasies and I have enjoyed our conversations immensely. I would then like to thank my co-supervisor Paul Grobler for his unending patience with me and the meaningful insights that he has given me. Then I would like to acknowledge the advice of one of my statistics lecturers, Sarel Steel, he has provided me with meaningful discussions and provides a reaffirming mirror for my ideas. Futhermore I would like to thank Kari du Plessis for her biological interpretation and effort applied to a portion of this work and the Computational Biology Group for listening to me rambling on about Bayes.

I would like to acknowledge the NRF for partial funding of this project.

Dedications

I would like to express my sincere gratitude to my family, my mother, father, sister, brother and friends, for without their support I would not be where I am today. I would also like to thank God for the insights and talents he has given me along with the inspiration to strive for new heights and the patience and tolerance to deal with the bumps in the road.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There has been a general increase in the amount of data that is generated in the field of genomics [6]. One of the drivers of this increase is the application of high-throughput technology, specifically in the field of sequencing [11]. This wealth of information has lead to an increased need for development in the field of functional genomics [1, 9, 8]. This may lead to an improvement of our functional understanding of the fundamental processes that occur within a cell by delineating the interactions of its fundamental components: genes. Therefore, this work will attempt to model the interaction from an exploratory approach in order to develop hypotheses about gene interactions that can then be further investigated. These hypotheses may potentially broaden our ability to manipulate and control organisms at a cellular level.

At the very basis of functional genomics is understanding the interaction between genes and gene products [5]. There are several possible ways to define what is meant by this interaction, and several ways to gain a conceptual understanding of such an interaction. Methods and approaches have been adapted from various fields in an attempt to approach the problem from different perspectives.

For the purposes of this thesis a more general definition of the concept of interaction is used. The interaction between genes is defined in terms of their relationship; this relationship can either be between their respective products, between their respective regulatory components or the genes themselves can have a regulatory relationship with each other. Given this context, we abstract the problem by visualizing it in terms of a network, or graph. Utilizing this context there are several possible ways to model the interaction between genes.

These methods can involve the application of a dynamical approach that attempts to model the problem based on assuming some nature for the interaction of these genes. This may be in the form of a system of differential equations, a set of binary interacting components or a network of interacting random variables [12, 13, 10]. Alternatively patterns prevalent in observed data can be uncovered, where the nature of these patterns may potentially indicate interactions. Patterns can be uncovered by modelling clusters of vari-

ables based on some measure of association. This association may take the form of similar observed responses, functional characteristics or variable similarity [3, 2].

In this thesis we aim to apply different approaches in an attempt to uncover the patterns in expression data, each based on certain assumptions. With the expression data putatively capturing the observed activity of genes under a set of perturbations. And the patterns uncovered from this data may allow for hypotheses on potential associations between genes.

Our first approach, a targetted analysis, aims to explore the relationship between a set of known variables. The association between a set of grapevine genes in the context of putative co-expression is investigated by taking into account dominant perturbations that may potentially drive the putative co-expression. The assumption underlying this approach is that co-expressed genes are assumed to have some common putative function.

We then aim to hypothesize relationships between putative metabolically related genes in an untargetted co-expression based approach. Here the assumption is made that sets of putatively co-expressed genes are influenced by other sets of putatively co-expressed genes across time. The relationships between genes are explored based on the statistical dependence between the observed expression of these putative sets in the context of yeast.

Alternatively, we also model the interaction between genes based on the assumption that subsets of co-occurring clustered genes act together, thus producing the observed expression values. We aim to hypothesize alternative relationships between these putative metabolically related genes. This is done using a combined statistical model to generate a network topology that is then summarized and investigated.

These methods are applied in the context of exploratory analysis and hypothesis generation with putative evidence provided in support. This evidence is gathered from current known biological information. Each of the methods highlighted above represents a novel approach in the attempt to capture putative gene interactions from patterns in gene expression data.

The thesis will first discuss the current literature, providing necessary information on the biological context and mathematical theory underlying the above methods, followed by a discussion on each respective exploratory method in the chapters that follow.

1.1 List of References

[1] Braga-Neto, U.M. and Marques Jr, E.T. (2006). From functional genomics to functional immunomics: new challenges, old problems, big rewards. *PLoS computational biology*, vol. 2, no. 7, p. e81.

[2] Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M., Ding, H., Xu, H., Han, J., Ingvarsdottir, K., Cheng, B., Andrews, B., Boone, C., Berger, S., Hieter, P., Zhang, Z., Brown, G., Ingles, J., Emili, A., Allis, C.D., Toczyksi, D.P., Weissman, J.S., Greenblatt, J. and Krogan, N. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, vol. 446, no. 7137, pp. 806–810.

[3] D'haeseleer, P., Liang, S. and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, vol. 16, no. 8, pp. 707–726.

[4] Hall, R.D., Brouwer, I.D. and Fitzgerald, M.A. (2008). Plant metabolomics and its potential application for human nutrition. *Physiologia plantarum*, vol. 132, no. 2, pp. 162–175.

[5] Hieter, P. and Boguski, M. (1997). Functional genomics: it's all how you read it. *Science*, vol. 278, no. 5338, pp. 601–602.

[6] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O. and Yon Rhee, S. (2008). Big data: The future of biocuration. *Nature*, vol. 455, no. 7209, pp. 47–50.

[7] Lu, P.Y., Xie, F. and Woodle, M.C. (2005). *In Vivo* application of rna interference: From functional genomics to therapeutics. *Advances in genetics*, vol. 54, pp. 115–142.

[8] Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, vol. 24, no. 3, pp. 133–141.

[9] Morozova, O. and Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, vol. 92, no. 5, pp. 255–264.

[10] Perrin, B.-E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J. and d'AlcheBuc, F. (2003). Gene networks inference using dynamic bayesian networks. *Bioinformatics*, vol. 19, no. suppl 2, pp. ii138–ii148.

[11] Reis-Filho, J.S. (2009). Next-generation sequencing. *Breast Cancer Res*, vol. 11, no. Suppl 3, p. 12.

[12] Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, vol. 18, no. 2, pp. 261–274.

[13] Tegner, J., Yeung, M.S., Hasty, J. and Collins, J.J. (2003). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5944–5949.

Chapter 2

Literature Review

To meaningfully discuss gene interaction networks and how to derive them, some preliminary information is needed, both in terms of biological background, mathematical theory and a general understanding of the current and prior developments in this field. The basic biology regarding gene interaction is discussed first and then an outline of microarray experiments is described. These experiments serve as the primary sources of data in the hypothesis generation approaches. A discussion of some of the important mathematical concepts used in this thesis as well as their related application in the literature completes this chapter.

Biological Context

With the increase in the availability of biological data, the potential meaningful available information of these systems also increases [58]. Currently, at the time of writing, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database consisted of 2,367 completed genomes and 9,592,536 genes [65]. The Gene Ontology contains 36,445 non-obsolete terms, used to annotate genes with their putative function and characterize them. The KEGG database has been in development since 1995, and is currently managed by Kanehisa Laboratories. It consists of the functional integrated information from genomic, systematic and chemical resources [65]. The Gene Ontology project consists of the description of gene products and it aims to provide a consistent description of biological processes and molecular functions [10]. These represent a significant proportion of the currently available knowledge regarding gene interaction. The network-based study of these biological systems have contributed significantly to this knowledge base [108, 26, 140].

The central dogma outlined in Figure 2.1 provides a summary of our global understanding of the fundamental interaction driving cellular function and development. Deoxyribonucleic acid (DNA) is a fundamental component of a cell. It consists of two polymer strings with repeating molecules called nucleotides, supported on a sugar phosphate backbone [137]. These nucleotides

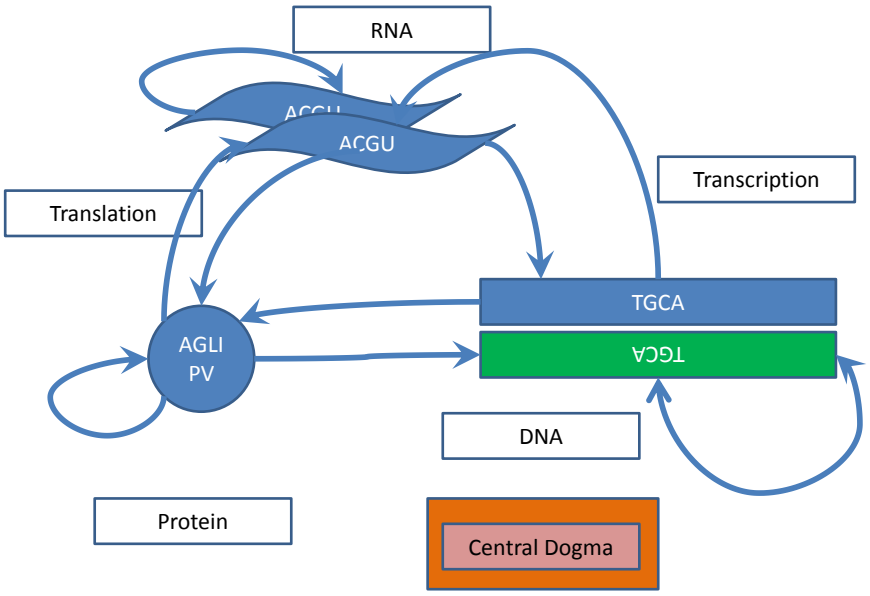


Figure 2.1: Central Dogma

The Central Dogma of molecular biology which illustrates the flow of information.

are often abstracted as four letters: A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). Together these form the building blocks for more complex structures in an organism. DNA is structured into chromosomes that both vary in number and size between species; as an example humans have 23 chromosomes that range in lengths from 51 million base pairs to 245 million base pairs. While on the other hand *Saccharomyces cerevisiae*, has only 16 chromosomes that vary in length between 230 000 and 1,530,000 base pairs. These chromosomes contain functional parts called genes, which encode the necessary information for cellular function. Genes are transcribed into ribonucleic acid (RNA), by the process of transcription, which is regulated by a protein called a transcription factor. These transcription factors are believed to bind to an upstream region of the gene, known as the promoter region.

Once a transcription factor binds to its binding region, RNA polymerase is then recruited. The DNA structure is then transcribed using the concept of complementary base pairing. This process is made more complex by many transcription factors potentially binding to the same or similar regions, and together they may activate or repress the transcription of a gene. The level of RNA produced by a specific gene is referred to as gene expression.

Here regulation refers to the interacting components that leads to the activation, repression or combination of both, of an individual or set of genes. A set of genes are said to have interacted if the products of these genes inter-

act, a subset of these genes are involved in the regulation of another subset of these genes or the set of genes regulate a set of genes outside this set. The information in messenger RNA (mRNA) that determines the protein sequence that will result from translation is encoded in sets of three contiguous nucleotides known as codons. These codons are then categorized into groups of amino acids which are the building block of proteins. Proteins as indicated in the depiction of central dogma, can, in turn, influence DNA, potentially influencing the transcriptional regulation of another protein or even proteins similar to itself, potentially repressing or activating transcription and eventually translation. The afore mentioned process effectively describes the central dogma and the situation becomes more involved as there are several additional factors that could influence this process. Essentially, gene regulation, which collectively refers to the various systems that impact gene expression, is a complex combinatorial dynamical system, which involves several interacting components [139, 107, 77, 121, 130].

Microarrays

One of the primary types of data that serves as a source for gene regulation studies in observing the expression of genes under various perturbations is microarray data. Microarray data measures the expression of thousands of genes at a time given a particular perturbation, or a collection of perturbations. The process involves the extraction of mRNA from a particular target cell. The messenger RNA, mRNA, is then reverse transcribed into complementary RNA, cRNA. To be able to monitor the expression levels of these cRNA's, they are labelled with fluorescent molecules. The cRNA is fragmented and then hybridized to a gene chip[110, 138]. With hybridization there is variable strength or regularity in the binding of nucleotide sequences to their complements. Depending on, amongst other things, the composition of the nucleotide sequence and the condition of the medium they reside in, some nucleotide sequences may bind more readily to others.[95].

This is an important bias that could inflate intensity values. Often the chip is also designed so that there are multiple probe sequences that match to a particular gene, thus probes are often referred to in terms of probesets. Gene chips consist of thousands of neatly organized spots and each spot has several probe sequences attached, where probes can be DNA or cDNA sequences. The chips are washed and then analysed using spectroscopy techniques to quantify the fluorescent signal produced at each spot. The quantitative results from each chip are then processed, resulting in a matrix of values. This matrix, called a gene expression matrix, has rows that indicate probe sequences and columns that indicate the chip and thus the perturbation applied. Often these rows represent probesets, rather than individual probes. This matrix is also the result of an extensive analysis process that attempts to reduce the technical and experimental noise of the entire process.

There are a number of concerns regarding gene expression data. Apart from the technical and experimental noise that may still remain and be propagated throughout the experimentation process, there is a distinct lack of degrees of freedom to consider. This occurs as there is generally a large number of variables measured with a very small number of samples. The experiments are also generally repeated only a few times and in some cases not repeated at all [124]. The main driver of this data problem is that microarray experiments are costly and they are also prone to experimental complications, the result of a noisy process, though there does exist evidence to contradict this popular belief [71]. Regardless, microarrays remain a popular source of information and several techniques and methods have been proposed in order to improve the information content of these data sources [117, 66].

There is a considerable amount of pre-processing that is done to obtain an expression matrix from a microarray experiment. The pre-processing could potentially reduce the influence that noise, both technical and biological, has on the data. Adjusting for systematic noise and bias may also allow for a more meaningful comparison of two different arrays. This process is referred to as normalization [111, 31]. One of the most popular methods for normalization of an Affymetrix GeneChip is Robust Multiarray Averaging (RMA) [61]. This process involves correcting for background noise, transforming the data by taking the log_2 of the intensity values and then applying quantile normalization [8].

It is important to note that this work is concerned with hypothesising gene interactions which is then depicted in a network structure. This means that the nodes of the network are genes, or as their respective proxies, probes or probesets. The concepts of nodes and networks will be explained further below.

Graph Theory

Identifying how the various components of these networks interact may improve our understanding of an organisms development or how these organism's response to various stresses and perturbations. Networks, or graphs, provide an intuitive way in which these interactions can be described, visualized and better understood. Not only has graph theory been used as a visualization tool for these complex networks, but also as a means to uncover properties that may have biological significance [51, 2].

The fundamental component in this work is a network, or graph, a concept formally explored by Euler in [34]. Here a graph, $\mathcal{G}$ is given by the sets $\{V, E\}$, with V a finite set referred to as the vertex set, also called a set of nodes, while E is called an edge set such that $(i, j) \in E, i \in V, j \in V$. Thus an edge, which is an element of the edge set, describes a connection between nodes. If the order of the connection is important, then the edge is called a directed edge, with the corresponding graph being referred to as a directed graph, alternatively both are referred to as undirected. In the case of an undirected graph, the elements

that constitute the edge are nodes, which are referred to as incident nodes to the edge. The number of edges that are incident to a node in an undirected graph is called the degree of the node. In a directed graph we have in-degree, the number of edges leading to the node, and out-degree, the number of edges leading from the node. The nodes that are connected by an edge are called adjacent to each other [14].

Together the all-against-all representation of adjacency can take the form of an adjacency matrix. Elements of this matrix can consist of binary values, where 1 indicates a relationship and 0 indicates the absence of one, alternatively any other value that indicates a weighted connection between nodes can be used to indicate the relationship between nodes. Here it is important to clarify that an adjacency matrix is a square matrix with dimension described by the vertex set, with an element in the matrix indicating a connection between the nodes corresponding to the respective row and column. This matrix is symmetrical when the graph is undirected and potentially asymmetric in the context of directed graphs. For clarity an example of an adjacency matrix for both the directed and undirected case is presented in Figure 2.2, the convention adopted here is that the source or parents of directed edges are indicated by the columns and the children or sink of a directed edge is given by the rows. Thus an edge from A to B is indicted by a 1 in the B, A entry of the adjacency matrix, assuming the weight of the connection is irrelevant or absent [14, 136].

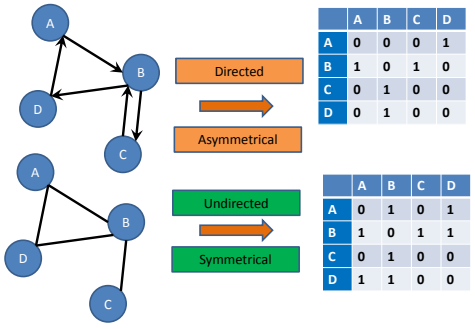


Figure 2.2: Adjacency Matrix Example

The figure illustrates an example of a directed graph (top), an undirected graph (bottom) and their corresponding adjacency matrices. As an example of adjacency in the undirected case, in the bottom figure nodes A and B are connected by an edge and are therefore adjacent, which corresponds to a 1 in the B,A entry as well as a 1 in the A,B entry of the corresponding adjacency matrix. In the directed case there is only a 1 in the B,A entry, indicating the edge goes from A to B.

In this thesis, gene interaction is at times investigated by considering the local structure of the network. This local structure is mathematically referred to as a subnetwork. With subnetworks defined in terms of networks as

Definition 2.1. (Subnetwork or Subgraph). A subnetwork, $\mathcal{G}^* = \{V^*, E^*\}$ is defined in terms of a graph $\mathcal{G} = \{V, E\}$, with $V^* \subseteq V$ and $E^* \subseteq E$, such that $E^* = \{(v_i, v_j) | \forall (v_i, v_j) \in E, v_i, v_j \in V^*\}$.

Thus putative gene interaction inference in terms of subgraphs is effectively a local structure learning problem. This problem can be approached in terms of inferring the relation between vertices in different clusters and/or vertices in the same cluster. The local inference of gene interaction can then be extracted for several subset of genes and combined to provide a putative approximation of the global gene interaction network.

Spanning trees are a particular type of subgraph. As a tree it contains no cycles and the spanning property ensures it contains all nodes of the original graph. Uncovering this particular subgraph from a graph structure is viewed as one of the first problems in graph theory, originally investigated by Euler in [34]. This is also considered to be the first formal published application of graph theory. If the edges in a graph have weights associated with them, then the problem of uncovering the spanning tree can be investigated in terms of finding a spanning tree of minimum weight. The history of approaches used to solve this problem is discussed in [53]. One of the first known algorithms for determining a minimum spanning tree is the one proposed by Borůvka in [15]. This algorithm is based on searching from one vertex to another and iteratively adding the edge of lowest weight until a tree is constructed. A criticism of the algorithm is that it does not take into account the historical edges added. Current algorithms used are Prim's algorithm, originally introduced in [62] and Kruskal's given in [73]

Various aspects of the network topology can be used to generate hypotheses on the biological nature of gene interactions. For a review on some of the biologically relevant aspects of network structure see [12]. Among the various topological characteristics of a network the concept of cliques and communities have found considerable interest in the study of genes [18].

A clique, C, is defined with regards to a graph G, as a complete subgraph. A graph is called complete if it contains all possible edges for its node set. A clique is said to be maximal if there does not exist another clique, H, such that C is contained within H. The problem of identifying the set of maximal cliques for a graph is closely related to the concept of finding the maximum clique in a graph. This problem is known to be NP-complete, thus no efficient time algorithm currently exists to solve it exactly, therefore approximate algorithms are generally applied [43]. In the application of clique detection the objective is often to identify dense structures or triangle type structures in a network. Depending on the node description these structures could have biological meaning. This concept has been applied to gene expression analysis, motif discovery, protein structure comparison and genetic variation disease studies related to single nucleotide polymorphisms [76, 11, 32, 75].

A related concept to cliques, that identify locally dense topologies, is that

of communities. A community is defined as a set of nodes for which edges within the community are more dense than edges between communities. Depending on the explicit definition used for a community, the problem is also NP-complete. Communities capture the intuition of highly interactive modules. Several algorithms have been proposed and applied in a variety of fields, for a comprehensive review of these algorithms see [38]. In the field of gene expression analysis, [126] argued for the application of networks in identifying gene co-expression modules to hypothesize functional relationships between sets of co-expressed genes by observing the resulting network topology. Recently, [129] applied community detection to a network of *Escherichia coli* putatively co-expressed genes. The application identified a robust, hierarchical, functionally related community structure, where communities were each significantly enriched in terms of Gene Ontology.

Functional Genomics

With the advent of high-throughput methods in the field of genomics, there has been a wealth of information generated. The area of functional genomics is concerned with the dynamical interaction between genes and gene products that can be derived from this information [106, 49]. One of the approaches used to functionally link genes is to compare their respective gene sequences. This is often done in phylogenetics, where the sequences are compared and aligned to hypothesize an evolutionary history of association [52]. The assumption is that genes that have similar phylogenetic profiles may be functionally related in some way [42, 104, 60].

Efficiently comparing large sequences is therefore an important aspect of functional genomics. Several methods to approximately match strings have been proposed, and a review of several algorithms is given in [64]. A popular method to compare sequences of nucleotides or amino acids is the Basic Local Alignment Search Tool (BLAST) [6]. This heuristic algorithm attempts to approximate a globally optimal alignment by maximizing a measure of local similarity. Alternative implementations of BLAST have been proposed to adapt the algorithm to more specific topical areas. ScalaBLAST optimizes BLAST comparison for larger datasets [100]. There are also BLAST implementations that are specifically used to compare specific sequence types such as nucleotide sequences with BLASTn, a translated nucleotide sequence to a set of protein sequences using BLASTx, a decoded protein sequence to a set of nucleotide sequences using tBLASTn and comparing protein sequences using BLASTp [7, 145, 19].

Dynamical Models

There have been several techniques developed over the years to both study and generate these networks. These techniques are quite varied and each

have their respective strengths and weaknesses. A comprehensive review of all the available techniques in the literature falls beyond the scope of this thesis, thus the set of models and techniques chosen provide an outline of the state of scientific knowledge with regards to the dynamical modelling of gene interaction networks.

Gene regulation can be viewed in a very simplified fashion as an interaction of 'on' or 'off' switches. One of the earlier models implemented by [51] utilized a type of boolean network to analyse the structure and behaviour of biochemical control networks. Boolean networks were also proposed by [74] and [3] in the study of gene regulation. This was extended by [119], [120] and [67] by adding a stochastic framework resulting in Probabilistic Boolean Networks, which incorporate certain elements of uncertainty. Boolean networks attempt to model the dynamics of gene interaction using rules in the form of boolean functions. In the model proposed by [67], nodes in the graph consist of genes, inputs and outputs, here each node takes on a binary value, edges indicate connection between the various processes and the states of the system evolve using Boolean Functions. These functions often take the form of logical operators such as AND, OR and XOR. The general limitation of these models is that they view expression of genes as binary, either 'on' or 'off' when it is known that gene regulation is more continuous. One of the benefits of these approaches is that they can uncover dynamical behaviour of gene regulatory networks, such as the study done by [78], where a yeast network was found to be resistant to small perturbations and the conclusion drawn that certain gene regulatory networks are robust.

The stochastic binary model of boolean networks can be extended to a more general continuous framework. The paper by [119] indicated the connection between the Probabilistic Boolean Networks and the Bayesian Network. The latter is another popular method for analysing gene regulatory networks. Bayesian Network models, in the study of gene regulation, have been proposed by, amongst others, [144], [40], [41] and [59]. These models are probabilistic in nature and benefit from their natural ability to incorporate prior information into the model. This allows the model to cope well with some of the inherent restrictions of microarray data. A Bayesian Network model attempts to model the interaction between nodes in a network as a product of conditional probability distributions. The set of variables are modelled in terms of a joint probability distribution that is factorized using conditional probabilities. The nature of this factorization is based on the dependence relationship between these variables. Bayesian Networks assume that these dependence relationships are given by the network structure itself. Therefore nodes are viewed as random variables that collectively form a joint probability distribution. The edge structure describes conditional dependence, thus the joint distribution can be factored into a product of conditional probability distributions. The application of Bayesian Networks is two-fold. Firstly, the structure itself can be viewed in a Bayesian context, where the structure can be estimated from a

posterior probability, viz.

$$P(G|D) \propto P(D|G)P(G),$$

here G is a random variable over graph structures with prior $P(G)$, D is the observed data with $P(D|G)$ the conditional probability of generating the data given the graph structure. The normalization constant is left out of the equation, as generally Markov Chain Monte Carlo simulations are used to sample graph structures from the posterior $P(G|D)$. Secondly, learning the actual conditional probabilities that govern the interactions can also be accomplished using a Bayesian framework.

A review of the reconstruction accuracy of a few of these Bayesian Network models is provided in [135], where the performance is evaluated on both synthetic and real data. In these networks, nodes are generally considered to be genes or gene products and the directed edges indicate putative connections between these components. A major drawback of Bayesian Networks with regards to gene regulation is that the network is required to be acyclic. This limits the ability of the model to capture feedback loops and auto-regulation. Bayesian Networks also lead to equivalent classes, because there may be several network structures that can explain the same probability distribution. In the comparative work of [135] it was argued that networks that can uncouple these equivalent classes show an improved performance. This uncoupling was done by either including intervention data acquired by data collection, additional information or knowledge of the system itself or by using Dynamic Bayesian Networks.

A Dynamic Bayesian Network (DBN) is a directed network that results from *unrolling* a Bayesian Network with respect to some dimension, often time. A toy example of this is given in Figure 2.3. These models have been applied to several systems, for example time series microarray data, in order to infer the edge structure of regulatory networks in *Saccharomyces Cerevisae*, [70]. They have also been applied as predictive models on a DNA repair network in *Escherichia coli* and [105], respectively. A further improved version of a DBN model was proposed by [149]. This model attempts to solve the computational burden of standard DBN models by using an initial variable selection phase. The results indicate an improvement in predictive accuracy of the gene regulatory network, with an improved overall run-time. Standard DBN models have considerably more variables than their Bayesian Network counterparts and the sample space of possible graph structure is thus much larger. This could potentially impact the effectiveness of structural inference, therefore there are often restrictions applied to the search space. The nature of these networks also allows for the modelling of feedback loops, which is an important known process in gene regulation.

The set of variables that have a statistical conditional dependence relationship with a target variable is often referred as the Markov Blanket of the

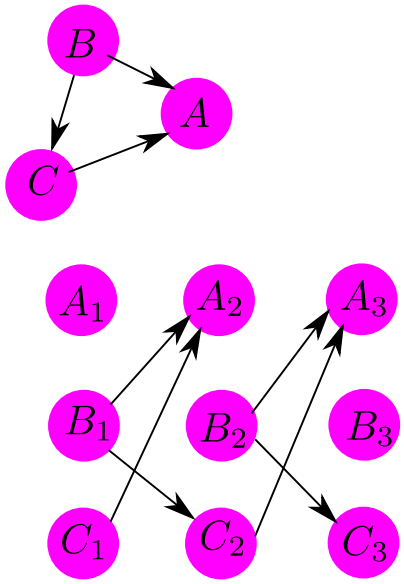


Figure 2.3: Dynamic Bayesian Network Example

The figure shows a directed network (top) unrolled over some index into a Dynamic Bayesian Network. Given some starting set of parameters, the edge structure of the unrolled network maintains the dependency on the original network. The network can be unrolled for arbitrary many indices. It is assumed here that the interaction defined occurs across indices and not within.

target variable.The term Markov Blanket was originally coined by [103]. In the context of a network these blankets describe the immediate neighbourhood of these target variables. When modelling gene interaction by statistical dependence, such as in Bayesian Network and Dynamic Bayesian Network models, we can essentially estimate the Markov Blankets for the respective variables [83]. There have been a number of applications of Markov Blankets and several algorithms developed to estimate them from data. A target variable's Markov Blanket is often described as the minimum set of variables for which information is needed in order to perform inference on this target variable. This concept has lead to the application of Markov Blankets to feature selection and set generation. In order to improve the robustness, computational efficiency and classification accuracy of genetic algorithms applied to gene expression data, [148] proposed the use of a Markov Blanket approach with comparative success. A review of the application of Markov Blankets to feature selection and inference of causal structure is given in [4]. In inferring network structure in the context of gene regulatory networks, [79] proposed the use of Markov Blanket detection combined with a set of ordinary differential equations, using local dynamics to infer the network structure. It was argued that the success of this method may depend highly on the Markov Blanket detection algorithm used as well as the Euler approximation applied to the differential equations

[99].

One of the most powerful methodologies that has been used to model the dynamics of these systems, and therefore to infer the nature of gene interaction, is a system of ordinary differential equations. A basic differential equation model was proposed by [22]. Their model was based on kinetic equations which incorporates degradation rates of mRNA and proteins. They also discussed two methods that could be used to infer the parameters of their model under different assumptions, however, their model was not applied to real experimental data. This model was extended by [29] and applied to data from *Bacillus subtilis*, their approach involved using Akaike's Information Criterion to infer the sparsity of the network rather than assuming a particular form of sparsity as done in *chen1999modeling*. Here sparsity is in reference to the number of genes which putatively influence the expression of a target gene, this number being relatively small. These models have several different forms, each attempting to model different aspects of gene interaction based on various assumptions. Two important limitations of these models are: they are often linear and therefore cannot fully capture the non-linear dynamics of the system and further, the models assume a static influence across the time-course. The latter results in the inability of the model to capture switch-like behaviour, whereby under certain perturbations the system might utilize different biological regulatory pathways. Alternative models have also been suggested, such as the system of differential equations proposed by [115], where parameters are estimated using genetic programming and least mean squares. Other examples of non-linear methods have also been proposed in [131], [133] and [45]. The latter model proposes the use of hill functions that are approximated using discrete stepwise functions and therefore applying a piecewise linear model to investigate the dynamics of the system. Differential equations are well studied and provide a detailed description of the dynamics underlying a system. The model depends highly on knowledge of the underlying system, which often involves a number of assumptions. These models also require large reliable datasets to estimate the parameters of the model. This has lead to a number of papers and discussion on artificial networks for benchmarking and parameter estimation. These artificial networks are generated using a system of differential equations or synthetic biological engineering, [118, 20]. It has also recently been shown that there exists a particular equivalence between the Euler approximation of a particular set of ordinary differential equations and a Dynamic Bayesian model based on a Gaussian Distribution and regression [99].

Artificial networks take the form of a hypothesized synthetic network, from which synthetic data is generated using the proposed system of differential equations. Thus if a model is applied to the synthetically generated data, the resultant inferred network can then be compared to the original hypothesized network. These types of networks are based on an assumed dynamic that govern the underlying system. Therefore datasets constructed from these networks, when utilized for exploratory analysis and hypothesis generation, may

be biased towards these assumed dynamics. The hypothesis proposed may simply reflect this dynamic instead of the real biological phenomena. Nevertheless these networks have been used extensively in algorithm development, providing a framework to benchmark methods in a machine learning context and assess their respective predictive capacities. These networks are not applied in this thesis as the focus is on exploratory analysis and hypothesis generation.

Metrics and Clustering

Apart from the models mentioned above, namely approaches that try to uncover the network structure by modelling the dynamics of the system, there are also methods that have been used to generate modules of putative interacting genes based on various metric and pattern recognition approaches. In particular, clustering along with a distance metric can give an indication of modules of genes that share some underlying characteristic. Modules of genes can refer to groups of genes that may have some regulatory relationship with respect to each other, may exhibit some common function in a cell or may even refer to genes that are part of the same biological pathway or process. The distance metric used may capture different perspectives of the relationship between genes and thus potentially uncover these putative relationships. These have been applied to identify groups of putatively co-expressed genes, under the assumption that these co-expressed genes have similar regulatory mechanisms or function [16, 17, 142]. In analysing gene expression matrices that are the results of microarray experiments, these approaches involve the application of various measures or metrics to vectors of intensity values. These then provide an indication of similarity of dissimilarity, which can then be used to cluster the respective vectors into meaningful sets. Measures that have been applied include Pearson Correlation, Spearman Correlation, Mutual Information and Euclidean Distance, amongst others [134, 96, 35, 146, 84].

For the analysis of gene expression data, Pearson correlation is one of the most widely used methods to model the gene expression patterns that are prevalent in the data. Rather than capturing the distance between gene expression vectors, potentially modelling the bias associated with hybridization kinetics, the shape of the vectors are compared. The variability of strength with which certain sequences hybridize compared to others may lead to comparatively higher intensity values for certain probe sequences or sets, which may bias the analysis if the objective is to model expression. It is important to note that correlation does not imply causation. However, it has been extensively used to infer an approximate dependence relationship between variables. This is especially prevalent in studies of genotype-phenotype relationships [116].

Given a particular method to capture the distance or similarity between vectors of gene expression data, clustering can then be performed. Techniques or methods to cluster gene expression data range from fast simple k-means, in which the data is partitioned into k sets, to more involved stochastic methods

such as those using Dirichlet Processes. A brief review of the complexity of gene expression data and clustering algorithms applied to these datasets is given in [63, 68]. A similar discussion in terms of network inference, highlighting the application of clustering and a more advanced dynamic model is given in [33].

Clustering can also result from the application of mixture models, as discussed in [86]. Mixture models are built on the fundamental assumption that the mixture or combination of multiple models, functions or distributions can be used to describe a particular phenomena. These models have been applied to gene expression (microarray data) with putatively meaningful results. In [87], expression data was successfully modelled by a mixture of Student-t Distributions and estimation of the number of components was performed using an Expectation Maximization Algorithm, with the overall objective of estimating clusters in a high-dimensional space.

A stochastic extension of mixture modelling for clustering is a Dirichlet Process Mixture, also explained in Chapter 5 and originally discussed in [9]. These models are an extension of the non-parametric Bayesian Dirichlet Process model [37]. They benefit from modelling an infinite number of parameters, therefore not assuming a fixed number of mixture components. The flexibility of these models allows for the modelling of diverse phenomena. The parameters of interest and clusters can be sampled from the posterior distribution with the aid of Markov Chain Monte Carlo methods [93]. Models that have been used to analyse gene expression data using Dirichlet Process Mixture have been proposed by [27, 30, 88] amongst others. These models differ in the sampling scheme used and how the results from the sampling scheme are used to indicate clusters.

Markov Chain Monte Carlo

The application of Bayesian models, such as the Bayesian Network and Dirichlet Process type models referred to earlier, often require the evaluation of complex functions. This complexity may simply make these evaluations computationally infeasible. This problem can be addressed using Monte Carlo methods or its extension, Markov Chain Monte Carlo. The concept of Monte Carlo methods was formally introduced by [92]. It involves the estimation of a quantity of interest related to the function by observing samples obtained from the particular function. The classical example is the calculation of the expectation of a function, say $g(x)$, that is related to some discrete probability distribution, say $\pi(x)$,

$$E_{\pi}[g(x)] = \sum_{x \in \omega} g(x)\pi(x), \tag{2.0.1}$$

in the discrete case, this involves the sum over the sample space ω. With the probability distribution defined for some random variable X, where x is a observation of the random variable X. If the summation is computationally

infeasible, the Monte Carlo estimation involves obtaining a random sample $x_1, x_2, x_3, \ldots x_n$ from the probability distribution. Then the empirical mean calculated from the result of evaluating the function at the sample values, can be used as an estimate for the expected value,

$$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(x_i). \tag{2.0.2}$$

This estimate can be justified by the Weak Law of Large Numbers, whereby the empirical mean of a sample approaches the true underlying mean as the sample size increases [69]. From this it can be shown that for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{g} - E_\pi[g(x)]| \geq \epsilon) = 0.$$

Depending on the functional form of $g(x)$, other quantities of interest can be estimated. If, for example, the objective is to estimate probabilities from $\pi(x)$ then the same procedure can be applied only with $g(x) = 1_s$, which is the indicator function. The indicator function takes on a value of 1 when $x = s$, and 0 otherwise. The use of Monte Carlo methods are often limited, as the fundamental assumption is that we can obtain a random sample from a target distribution $\pi(x)$. Bayesian methods often lead to complex distributions, for which direct sampling may be computationally intractable.

Markov Chain Monte Carlo methods, originally introduced in [91], can be applied to sample from complicated target distributions and involve sampling from a constructed Markov chain [50]. Let us define a Markov Chain,

Definition 2.2. (Markov Chain). Given a discrete state space, s, and discrete time space, t. A Markov Chain is given by

$$\{X_t = s; t = 0, 1, 2, 3, \ldots, s = 0, 1, 2, 3, \ldots\}, \tag{2.0.3}$$

such that

$$P(X_{t+1} = s_{t+1}|X_t = s_t, X_{t-1} = s_{t=1}, \ldots, X_0 = s_0) = P(X_{t+1} = s_{t+1}|X_t = s_t) \tag{2.0.4}$$

where $P_{ij} = P(X_{t+1} = j|X_t = i)$ denotes the transition probability from state i to state j, where the transition probabilities given here are assumed to be the same for all time points. The matrix of all transition probabilities is therefore denoted as P.

The Markov Chain is constructed in such a way that the limit distribution of the chain is the target distribution $\pi(x)$. These chains are completely defined by their respective transition probabilities, P, thus effectively the process involves determining appropriate transition probabilities to construct the chain. Transition probabilities define the probability of transitioning from one

state to another, where a state is an element from the support of the target distribution. If the transition probabilities are invariant with respect to time, then the chain is homogeneous. The assumption here is that we have a discrete state space and discrete time space, this however can be generalized to their respective continuous counterparts where we then refer to transition kernels and Markov Processes. We further restrict our discussion to homogeneous Markov Chains.

There are a few essential properties that are required of the chain in order to ensure that the samples generated from the chain are from the target distribution [48]. Firstly, a sufficient condition for some limiting distribution to be our target distribution. This is achieved if the chain satisfies *general balance* with respect to our probability vector $\pi(x)$,

$$\pi(x)P = \pi(x). \tag{2.0.5}$$

Here $\pi(x)$ is called the stationary distribution for the transition probabilities and indicates a row vector of probabilities concerned with the random variable X. Secondly, to ensure that the chain converges to a distribution, the chain must be *ergodic*. For a chain to be *ergodic*, it needs to be *irreducible* and *aperiodic*. For an *irreducible* chain we require that from any given state there is a positive probability of reaching any other state. An *aperiodic* chain implies there is no fixed interval of time steps, greater than 1, for which the chain always returns to its current state. With the *ergodic* property we ensure that the chain converges to some distribution, the sufficient condition for this distribution to be our target distribution is given by *general balance*. The property of *general balance* is hard to verify as it may require, in the discrete case, a summation over a complicated state space, where the *general balance* criteria above, implies

$$\forall x^* \in S, \sum_{x \in S} \pi(x)P(x \to x^*) = \pi(x^*). \tag{2.0.6}$$

To overcome this we can require that our chain satisfies the more stringent property of *detailed balance*. This is given by,

$$\forall x, x^* \in S; \pi(x)P(x \to x^*) = \pi(x^*)P(x^* \to x), \tag{2.0.7}$$

where $P(x \to x^*)$ refers to the transition probability from state x to state x^*.

A means of constructing a transition probability matrix that satisfies the above criteria was provided by [55]. This is achieved by defining the off-diagonal transition probabilities by

$$P(x \to x^*) = Q(x \to x^*)A(x \to x^*), \tag{2.0.8}$$

where Q is any transition probability matrix, also referred to as the proposal distribution and A is defined by,

$$A(x \to x^*) = min\{1, \frac{\pi(x^*)Q(x \to x^*)}{\pi(x)Q(x^* \to x)}\} \tag{2.0.9}$$

with the restriction that $\forall x, x^* \in S; A(x \to x^*) = 0$ *if* $Q(x \to x^*) = 0$, A is often also referred to as the acceptance probability of the proposed value. The diagonal elements are defined such that the respective row sums of P are equal to one. Say, $\pi(x)$, our target distribution, is known only upto some proportion,

$$\pi(x) = \frac{h(x)}{c}, \tag{2.0.10}$$

where we know h and c is some, often intractable, normalization constant. Then 2.0.9 becomes

$$A(x \to x^*) = min\{1, \frac{h(x^*)Q(x \to x^*)}{h(x)Q(x^* \to x)}\}, \tag{2.0.11}$$

where we have that the normalization constants cancel out. This construction of transition probabilities by using the above ensures that we have *detailed balance*. All that remains when constructing these chains is to ensure that our transition probabilities result in a chain that is *ergodic*. We can construct *irreducible* chains by combining transition probability matrices that are not *irreducible* but satisfy *detailed balance*. If a set of transition matrices satisfy *detailed balance*, $\{P_j; j = 1, 2, \dots n\}$, then we have that

$$P = P_1 P_2 \dots P_n \tag{2.0.12}$$

and

$$P = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{2.0.13}$$

also satisfy *detailed balance*. We can use 2.0.12, to construct a MCMC scheme that is defined for each random variable separately.

When sampling from a MCMC scheme it is important to consider the *burn-in* period, *thinning* and *convergence*. The Markov Chain will asymptotically approach the target distribution, *convergence* is concerned with how long we need to iterate the chain for until we are assured that we are sampling from the target distribution. There is no explicit way to determine if a chain has converged, there are however heuristics that can be used to give an indication if the chain has not yet converged. These include visualization methods and test statistics. Visualization methods comprise the use of autocorrelation plots and trace plots. Trace plots that do not contain large shifts in the sampled value putatively indicate that we may have convergence. By the very nature of

Markov Chains, the samples produced are autocorrelated, plots that indicate an exponential decay in the autocorrelation for larger lag values are putatively indicative of convergence.

Test statistics such as Geweke [47], which produces a z-score involves calculating the difference between the the mean of the first $n_{initial}$ proportion of sampled values and the mean of the last n_{final} proportion of sampled values, normalized by the asymptotic standard error. The normalized differences asymptotically follow a standard normal distribution, thus producing a z-score. This approach is discussed in [25]. One of the major criticisms in the application of this test statistic is the unclear choice of $n_{initial}$ and n_{final}, though a value of $n_{initial} = 0.1$ and $n_{final} = 0.5$ was originally suggested. Raftery and Lewis also discussed diagnostic strategy to determine the number of iterations required to achieve a certain level of precision [112]. A diagnostic test developed by Gelman and Rubin, involves the comparison of two separate sample runs, thus producing separate chains. The variances between chains are compared to the variances within chains. A resulting ratio value far from one therefore provides evidence to suggest a lack of convergence [46].

Gene expression data, specifically microarray data, remains one of the primary data sources for the study of gene interaction, gene expression and regulatory inference. New techniques and approaches are constantly being developed from various fields of study. One of the potential reasons for the development of such a variety of techniques, apart from the fact that they may offer different perspectives or exploit different aspects of the system, is that validation of these techniques is problematic [118]. As mentioned, previously, there are many techniques based on synthetic benchmarking. One of the major reasons for this is that the true underlying gene interaction network is mostly unknown, therefore biologically truthful validation is often complicated. Nevertheless, the complex dynamic of the biological system can often not be captured by hypothesized synthetic data, thus it is important to generate putative biologically relevant hypotheses that can be tested and evaluated using other experimental techniques.

The uncovering, learning or reverse engineering of the interaction between genes from available data has been, and still remains, a complicated problem. Viewing the problem from the point of view of a dynamical system and network reconstruction has lead to the development of several algorithms and methods. Each method seeks to model an aspect of the biological system, but the success of these methods are quite varied with no true benchmark currently available. With the increase in availability of data, so also the potential to uncover true interaction between genes increases. It is evident from the above that there has been, and still is, considerable effort committed to trying to understand the interaction between genes. Effort both from a computational mathematical perspective and from an experimental biological focus. The problem is further complicated by the combinatorial nature of gene interactions, the multiple levels of regulation that can effect gene expression and that genes, along

with their products, are part of a larger complex system, with several unknown components. Still, the techniques that have been applied have enjoyed various degrees of success and each provide another step in the model generation procedure, thus improving the general understanding of the system. The relevance of understanding how genes interact with each other lies in the fact that genes are fundamental functional units that play an important role in the cell of an organism.

2.1 List of References

[1] Äijö, T. and Lähdesmäki, H. (2009). Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, vol. 25, no. 22, pp. 2937–2944.

[2] Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, vol. 7, no. 3, pp. 243–255.

[3] Akutsu, T., Miyano, S. and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In: *Pacific Symposium on Biocomputing*, vol. 4, pp. 17–28. World Scientific Maui, Hawaii.

[4] Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S. and Koutsoukos, X.D. (2010). Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *The Journal of Machine Learning Research*, vol. 11, pp. 171–234.

[5] Aliferis, C.F., Tsamardinos, I. and Statnikov, A. (2003). Hiton: a novel markov blanket algorithm for optimal variable selection. In: *AMIA Annual Symposium Proceedings*, vol. 2003, p. 21. American Medical Informatics Association.

[6] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410.

[7] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402.

[8] Amaratunga, D. and Cabrera, J. (2001). Analysis of data from viral dna microchips. *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1161–1170.

[9] Antoniak, C.E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, pp. 1152–1174.

[10] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J.T. Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, vol. 25, no. 1, p. 25.

[11] Baldwin, N.E., Collins, R.L., Langston, M.A., Symons, C.T., Leuze, M.R. and Voy, B.H. (2004). High performance computational tools for motif discovery. In: *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, p. 192. IEEE.

[12] Barabási, A.-L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113.

[13] Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, vol. 6, no. 3-4, pp. 281–297.

[14] Bondy, J.A. and Murty, U.S.R. (1976). *Graph theory with applications*, vol. 290. Macmillan London.

[15] Borůvka, O. (1926). O jistém problému minimálním [on a certain minimal problem].

[16] Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS letters*, vol. 480, no. 1, pp. 17–24.

[17] Brohée, S., Janky, R., Abdel-Sater, F., Vanderstocken, G., André, B. and van Helden, J. (2011). Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic acids research*, vol. 39, no. 15, pp. 6340–6358.

[18] Butenko, S. and Wilhelm, W.E. (2006). Clique-detection models in computational biochemistry and genomics. *European Journal of Operational Research*, vol. 173, no. 1, pp. 1–17.

[19] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. (2009). Blast+: architecture and applications. *BMC bioinformatics*, vol. 10, no. 1, p. 421.

[20] Cantone, I., Marucci, L., Iorio, F., Ricci, M., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D. and Cosma, M. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, vol. 137, no. 1, p. 172.

[21] Chaouiya, C. (2007). Petri net modelling of biological networks. *Briefings In Bioinformatics*, vol. 8, no. 4, pp. 210–219.

[22] Chen, T., He, H. and Church, G. (1999). Modeling gene expression with differential equations. In: *Pacific symposium on biocomputing*, vol. 4, p. 4.

[23] Chickering, D. (1996). Learning bayesian networks is np-complete. *Lecture notes in statistics-New York-Springer Verlag-*, pp. 121–130.

[24] Covert, M., Schilling, C. and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of theoretical biology*, vol. 213, no. 1, pp. 73–88.

[25] Cowles, M.K. and Carlin, B.P. (1996). Markov chain monte carlo convergence diagnostics comparative review. *Journal of American Statistical Association*, vol. 91, pp. 883–904.

[26] Csermely, P., Agoston, V. and Pongor, S. (2004). The efficiency of multi-target drugs: the network approach might help drug design. *arXiv preprint q-bio/0412045*.

[27] Dahl, D. (2006). Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pp. 201–218.

[28] Davidson, E., Rast, J., Oliveri, P., Ransick, A., Calestani, C., Yuh, C., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, Alistair G. Pan, Z.j., Schilstra, M.J., Clarke, P.J., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L. and Bolouri, H. (2002). A genomic regulatory network for development. *Science Signalling*, vol. 295, no. 5560, p. 1669.

[29] De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N. and Miyano, S. (2002). Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. In: *Biocomputing 2003: Proc. Pacific Symposium*, vol. 8, pp. 17–28.

[30] Do, K.-A., Müller, P. and Tang, F. (2005). A bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 3, pp. 627–644.

[31] Dudoit, S. and Yang, J.Y.H. (2003). Bioconductor r packages for exploratory analysis and normalization of cdna microarray data. In: *The Analysis of Gene Expression Data*, pp. 73–101. Springer.

[32] Dukka, B., Akutsu, T., Tomita, E., Seki, T. and Fujiyama, A. (2002). Point matching under non-uniform distortions and protein side chain packing based on an efficient maximum clique algorithm. In: *Genome informatics. International Conference on Genome Informatics*, vol. 13, p. 143.

[33] DâĂŹhaeseleer, P., Liang, S. and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, vol. 16, no. 8, pp. 707–726.

[34] Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, vol. 8, pp. 128–140.

[35] Ewing, R., Kahla, A., Poirot, O., Lopez, F., Audic, S. and Claverie, J. (1999). Large-scale statistical analyses of rice ests reveal correlated patterns of gene expression. *Genome research*, vol. 9, no. 10, pp. 950–959.

[36] FANG-XIANG, W., Zhang, W. and ANTHONY, J. (2004). State-space model with time delays for gene regulatory networks. *Journal of biological Systems*, vol. 12, no. 04, pp. 483–500.

[37] Ferguson, T.S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pp. 209–230.

[38] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, vol. 486, no. 3, pp. 75–174.

[39] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science Signalling*, vol. 303, no. 5659, p. 799.

[40] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620.

[41] Friedman, N., Nachman, I. and Peér, D. (1999). Learning bayesian network structure from massive datasets: the «sparse candidate «algorithm. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 206–215. Morgan Kaufmann Publishers Inc.

[42] Galperin, M.Y. and Koonin, E.V. (2000). Who's your neighbor? new computational approaches for functional genomics. *Nature biotechnology*, vol. 18, no. 6, pp. 609–613.

[43] Gary, M.R. and Johnson, D.S. (1979). Computers and intractability: A guide to the theory of np-completeness.

[44] Gat-Viks, I., Tanay, A., Raijman, D. and Shamir, R. (2006). A probabilistic methodology for integrating knowledge and experiments on biological networks. *Journal of Computational Biology*, vol. 13, no. 2, pp. 165–181.

[45] Gebert, J., Radde, N. and Weber, G. (2007). Modeling gene regulatory networks with piecewise linear differential equations. *European journal of operational research*, vol. 181, no. 3, pp. 1148–1165.

[46] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pp. 457–472.

[47] Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Department.

[48] Geyer, C. (2011). Introduction to markov chain monte carlo. *Handbook of Markov Chain Monte Carlo*, pp. 3–48.

[49] Gibson, G. and Muse, S.V. (2002). *A primer of genome science*. Sinauer Sunderland^ eMass Mass.

[50] Gilks, W., Richardson, S. and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in practice: interdisciplinary statistics*, vol. 2. Chapman & Hall/CRC.

[51] Glass, L. and Kauffman, S. (1973). The logical analysis of continuous, nonlinear biochemical control networks. *Journal of Theoretical Biology*, vol. 39, no. 1, pp. 103–129.

[52] Gould, S.J. (1977). *Ontogeny and phylogeny*. Harvard University Press.

[53] Graham, R.L. and Hell, P. (1985). On the history of the minimum spanning tree problem. *Annals of the History of Computing*, vol. 7, no. 1, pp. 43–57.

[54] Guet, C., Elowitz, M., Hsing, W. and Leibler, S. (2002). Combinatorial synthesis of genetic networks. *Science*, vol. 296, no. 5572, pp. 1466–1470.

[55] Hastings, W.K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, vol. 57, no. 1, pp. 97–109.

[56] Heckerman, D., Geiger, D. and Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, vol. 20, no. 3, pp. 197–243.

[57] Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T., Charnock-Jones, D. and Miyano, S. (2008). Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, vol. 24, no. 7, pp. 932–942.

[58] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O. and Yon Rhee, S. (2008). Big data: The future of biocuration. *Nature*, vol. 455, no. 7209, pp. 47–50.

[59] Husmeier, D., Dybowski, R. and Roberts, S. (2004). *Probabilistic modeling in bioinformatics and medical informatics*. Springer.

[60] Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research*, vol. 10, no. 8, pp. 1204–1210.

[61] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, vol. 4, no. 2, pp. 249–264.

[62] Jarník, V. (1930). O jistém problému minimálním. *Práca Moravské Prírodovedecké Spolecnosti*, vol. 6, pp. 57–63.

[63] Jiang, D., Tang, C. and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 11, pp. 1370–1386.

[64] Jokinen, P., Tarhio, J. and Ukkonen, E. (1996). A comparison of approximate string matching algorithms. *Software: Practice and Experience*, vol. 26, no. 12, pp. 1439–1458.

[65] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, vol. 40, no. D1, pp. D109–D114.

[66] Kathleen Kerr, M. and A CHURCHILL, G. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical research*, vol. 77, no. 02, pp. 123–128.

[67] Kauffman, S., Peterson, C., Samuelsson, B. and Troein, C. (2003). Random boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences*, vol. 100, no. 25, pp. 14796–14799.

[68] Kerr, G., Ruskin, H.J., Crane, M. and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, vol. 38, no. 3, pp. 283–293.

[69] Khintchine, A. (1929). Sur la loi des grands nombres. *Comtes rendus de l'Académie des sciences*, vol. 188, pp. 477–479.

[70] Kim, S., Imoto, S. and Miyano, S. (2003). Inferring gene networks from time series microarray data using dynamic bayesian networks. *Briefings in bioinformatics*, vol. 4, no. 3, pp. 228–235.

[71] Klebanov, L. and Yakovlev, A. (2007). How high is the level of technical noise in microarray data. *Biol Direct*, vol. 2, no. 9.

[72] Koch, I., Schuler, M. and Heiner, M. (2005). Stepp-search tool for exploration of petri net paths: a new tool for petri net-based path analysis in biochemical networks. *In silico biology*, vol. 5, no. 2, pp. 129–138.

[73] Kruskal, J.B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50.

[74] Lähdesmäki, H., Shmulevich, I. and Yli-Harja, O. (2003). On learning gene regulatory networks under the boolean network model. *Machine Learning*, vol. 52, no. 1, pp. 147–167.

[75] Lancia, G., Bafna, V., Istrail, S., Lippert, R. and Schwartz, R. (2001). Snps problems, complexity, and algorithms. In: *AlgorithmsâĂŤESA 2001*, pp. 182–193. Springer.

[76] Langston, M.A., Lin, L., Peng, X., Baldwin, N.E., Symons, C.T., Zhang, B. and Snoddy, J.R. (2005). A combinatorial approach to the analysis of differential gene expression data. In: *Methods of Microarray Data Analysis*, pp. 223–238. Springer.

[77] Le Hir, H., Nott, A. and Moore, M. (2003). How introns influence and enhance eukaryotic gene expression. *Trends in biochemical sciences*, vol. 28, no. 4, pp. 215–220.

[78] Li, F., Long, T., Lu, Y., Ouyang, Q. and Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 14, pp. 4781–4786.

[79] Li, Z., Li, P., Krishnan, A. and Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, vol. 27, no. 19, pp. 2686–2691.

[80] Li, Z., Shaw, S., Yedwabnick, M. and Chan, C. (2006). Using a state-space model with hidden variables to infer transcription factor activities. *Bioinformatics*, vol. 22, no. 6, pp. 747–754.

[81] Lu, T., Khalil, A. and Collins, J. (2009). Next-generation synthetic gene networks. *Nature biotechnology*, vol. 27, no. 12, pp. 1139–1150.

[82] MacKay, D. (1998). Introduction to monte carlo methods. *Nato Asi Series D Behavioural and Social Sciences*, vol. 89, pp. 175–204.

[83] Margaritis, D. and Thrun, S. (1999). Bayesian network induction via local neighborhoods. Tech. Rep., DTIC Document.

[84] Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, vol. 7, no. Suppl 1, p. S7.

[85] McAdams, H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, vol. 94, no. 3, pp. 814–819.

[86] McLachlan, G.J. and Basford, K.E. (1988). Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, vol. 1.

[87] McLachlan, G.J., Bean, R. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, vol. 18, no. 3, pp. 413–422.

[88] Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206.

[89] Mendes, P. (1999). Metabolic simulation as an aid in understanding gene expression data.

[90] Mestl, T., Lemay, C. and Glass, L. (1996). Chaos in high-dimensional neural and gene networks. *Physica D: Nonlinear Phenomena*, vol. 98, no. 1, pp. 33–52.

[91] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, vol. 21, p. 1087.

[92] Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, vol. 44, no. 247, pp. 335–341.

[93] Neal, R.M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265.

[94] Ness, S. (2007). Microarray analysis: basic strategies for successful experiments. *Molecular biotechnology*, vol. 36, no. 3, pp. 205–219.

[95] Nguyen, D., Bulak Arpat, A., Wang, N. and Carroll, R. (2004). Dna microarray experiments: biological and technological aspects. *Biometrics*, vol. 58, no. 4, pp. 701–717.

[96] NINDS, N. (1998). Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data. In: *Information processing in cells and tissues*, p. 203. Plenum Pub Corp.

[97] Nonrestrictive, I. (1999). Fundamentally different logic minireview of gene regulation in eukaryotes and prokaryotes. *Cell*, vol. 98, pp. 1–4.

[98] Novak, B. and Tyson, J. (1997). Modeling the control of dna replication in fission yeast. *Proceedings of the National Academy of Sciences*, vol. 94, no. 17, pp. 9147–9152.

[99] Oates, C., Hill, S. and Mukherjee, S. (2012). On the relationship between odes and dbns. *arXiv preprint arXiv:1201.3380*.

[100] Oehmen, C. and Nieplocha, J. (2006). Scalablast: A scalable implementation of blast for high-performance data-intensive bioinformatics analysis. *Parallel and Distributed Systems, IEEE Transactions on*, vol. 17, no. 8, pp. 740–749.

[101] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, vol. 27, no. 1, pp. 29–34.

[102] Palsson, B. (2006). *Systems biology: properties of reconstructed networks*. Cambridge University Press.

[103] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

[104] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, vol. 96, no. 8, pp. 4285–4288.

[105] Perrin, B., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J. and dÂŠAlche Buc, F. (2003). Gene networks inference using dynamic bayesian networks. *Bioinformatics*, vol. 19, no. suppl 2, pp. ii138–ii148.

[106] Pevsner, J. (2009). *Bioinformatics and functional genomics*. John Wiley & Sons.

[107] Proudfoot, N. (1986). Transcriptional interference and termination between duplicated α-globin gene constructs suggests a novel mechanism for gene regulation.

[108] Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. and Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1053–1066.

[109] Quach, M., Brunel, N. and d'Alché Buc, F. (2007). Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, vol. 23, no. 23, pp. 3209–3216.

[110] Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, vol. 2, no. 6, pp. 418–427.

[111] Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, vol. 32, pp. 496–501.

[112] Raftery, A.E. and Lewis, S. (1992). How many iterations in the gibbs sampler. *Bayesian statistics*, vol. 4, no. 2, pp. 763–773.

[113] Ribeiro, A., Zhu, R. and Kauffman, S. (2006). A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology*, vol. 13, no. 9, pp. 1630–1639.

[114] Richard, A. and Comet, J. (2007). Necessary conditions for multistationarity in discrete dynamical systems. *Discrete Applied Mathematics*, vol. 155, no. 18, pp. 2403–2413.

[115] Sakamoto, E. and Iba, H. (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming. In: *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, pp. 720–726. IEEE.

[116] Salvatore, F., Scudiero, O. and Castaldo, G. (2002). Genotype–phenotype correlation in cystic fibrosis: the role of modifier genes. *American Journal of Medical Genetics*, vol. 111, no. 1, pp. 88–95.

[117] Sanguinetti, G., Milo, M., Rattray, M. and Lawrence, N. (2005). Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, vol. 21, no. 19, pp. 3748–3754.

[118] Schaffter, T., Marbach, D. and Floreano, D. (2011). Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, vol. 27, no. 16, pp. 2263–2270.

[119] Shmulevich, I., Dougherty, E., Kim, S. and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, vol. 18, no. 2, pp. 261–274.

[120] Shmulevich, I., Gluhovsky, I., Hashimoto, R., Dougherty, E. and Zhang, W. (2003). Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks. *Comparative and functional genomics*, vol. 4, no. 6, pp. 601–608.

[121] Smale, S. (2001). Core promoters: active contributors to combinatorial gene regulation. *Genes & development*, vol. 15, no. 19, pp. 2503–2508.

[122] Smith, J., Theodoris, C. and Davidson, E. (2007). A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science Signalling*, vol. 318, no. 5851, p. 794.

[123] Smolen, P., Baxter, D. and Byrne, J. (1999). Effects of macromolecular transport and stochastic fluctuations on dynamics of genetic regulatory systems. *American Journal of Physiology-Cell Physiology*, vol. 277, no. 4, pp. C777–C790.

[124] Smyth, G., Yang, Y. and Speed, T. (2003). Statistical issues in cdna microarray data analysis. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-*, vol. 224, pp. 111–136.

[125] Stolovitzky, G., Monroe, D. and Califano, A. (2007). Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences*, vol. 1115, no. 1, pp. 1–22.

[126] Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, vol. 302, no. 5643, pp. 249–255.

[127] Thomas, R. (1996). Feedback loops: the wheels of regulatory networks. *Integrative Approaches to Molecular Biology*, pp. 167–178.

[128] Thomas, R., Thieffry, D. and Kaufman, M. (1995). Dynamical behaviour of biological regulatory networksÂŮi. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bulletin of mathematical biology*, vol. 57, no. 2, pp. 247–276.

[129] Treviño III, S., Sun, Y., Cooper, T.F. and Bassler, K.E. (2012). Robust detection of hierarchical communities from escherichia coli gene expression data. *PLoS Computational Biology*, vol. 8, no. 2, p. e1002391.

[130] Tuch, B., Galgoczy, D., Hernday, A., Li, H. and Johnson, A. (2008). The evolution of combinatorial gene regulation in fungi. *PLoS biology*, vol. 6, no. 2, p. e38.

[131] Tyson, J., Csikasz-Nagy, A. and Novak, B. (2002). The dynamics of cell cycle regulation. *Bioessays*, vol. 24, no. 12, pp. 1095–1109.

[132] Van den Bulcke, T., Van Leemput, K., Naudts, B., Van Remortel, P., Ma, H., Verschoren, A., De Moor, B. and Marchal, K. (2006). Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, vol. 7, no. 1, p. 43.

[133] Vu, T. and Vohradsky, J. (2007). Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of *Saccharomyces cerevisiae*. *Nucleic acids research*, vol. 35, no. 1, pp. 279–287.

[134] Wen, X., Fuhrman, S., Michaels, G., Carr, D., Smith, S., Barker, J. and Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences*, vol. 95, no. 1, pp. 334–339.

[135] Werhli, A. (2012). Comparing the reconstruction of regulatory pathways with distinct bayesian networks inference methods. *BMC Genomics*, vol. 13, no. Suppl 5, p. S2.

[136] West, D.B. (2001). *Introduction to graph theory*, vol. 2. Prentice hall Englewood Cliffs.

[137] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Genes: Structure, replication and expression. In: *Prescott's microbiology*, chap. 12. McGraw-Hill.

[138] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Microbial genomics. In: *Prescott's microbiology*, chap. 16. McGraw-Hill.

[139] Wu, L. and Belasco, J. (2008). Let me count the ways: mechanisms of gene regulation by mirnas and sirnas. *Molecular cell*, vol. 29, no. 1, pp. 1–7.

[140] Xu, J. and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805.

[141] Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, T. and Miyano, S. (2007). Finding module-based gene networks with state-space models-mining high-dimensional and short time-course gene expression data. *Signal Processing Magazine, IEEE*, vol. 24, no. 1, pp. 37–46.

[142] Yang, E., Foteinou, P., King, K., Yarmush, M. and Androulakis, I. (2007). A novel non-overlapping bi-clustering algorithm for network generation using living cell array data. *Bioinformatics*, vol. 23, no. 17, pp. 2306–2313.

[143] Yeung, M., Tegnér, J. and Collins, J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 6163–6168.

[144] Yoo, C. and Cooper, G. (2002). Discovery of gene-regulation pathways using local causal search. In: *Proceedings of the AMIA Symposium*, p. 914. American Medical Informatics Association.

[145] Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000). A greedy algorithm for aligning dna sequences. *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 203–214.

[146] Zhou, X., Wang, X. and Dougherty, E. (2003). Construction of genomic networks using mutual-information clustering and reversible-jump markov-chain-monte-carlo predictor design. *Signal Processing*, vol. 83, no. 4, pp. 745–761.

[147] Zhu, R., Ribeiro, A., Salahub, D. and Kauffman, S. (2007). Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. *Journal of theoretical biology*, vol. 246, no. 4, p. 725.

[148] Zhu, Z., Ong, Y.-S. and Dash, M. (2007). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, vol. 40, no. 11, pp. 3236–3248.

[149] Zou, M. and Conzen, S. (2005). A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, vol. 21, no. 1, pp. 71–79.

Chapter 3

Targetted Co-Expression Analysis

3.1 Introduction

Proteins are fundamental components of a living organism. They play an essential role in governing cellular function and development; where these proteins consist of one or more polypeptide chains of amino acids. Central Dogma maintains that these polypeptide chains of amino acids are the result of covalently linking decoded mRNA, a process known as translation [40]. The association between genes and proteins is not necessarily one to one. Proteins are often the result of a combinatorial interaction of genes, whereby genes are said to interact if their respective products interact with eachother, they have some regulatory role with respect to each other or they have some common regulatory mechanism. This complex dynamical interaction is one of the fundamental processes that influence the complex mechanisms of a cell. Understanding the functional association of these genes is therefore vital to delineating the nature of their interaction and how this information is propagated to the rest of the cell. As a gene can be involved in many processes, it also may have more than one function [41].

A classical approach to inferring functions of genes, known as genome annotation, involves determining the function of similar genes. Here similarity can refer to similarity in terms of sequences (nucleotide or amino acid). Putative functions for genes can be thus be assigned by using alignment and phylogenetic techniques to relate a particular gene to some gene with a putative function. The fundamental assumption underlying such an approach is that genes that have similar sequences have similar functions [41].

An alternative, or complementary method, is to study genes in sets. This may have an advantage over the study of genes in isolation, which may miss important patterns and information that is prevalent in the data. Evidence for the validity of this approach is gained by the knowledge that multiple genes interact with each other and together they potentially influence several processes [41]. Modules of genes can be produced from microarray data by

identifying putatively co-expressed sets of genes, based on their estimated correlation. These sets can then be analysed to infer potential biological functions of the genes in the set. The assumption here is that genes that are putatively co-expressed under a particular biological perturbation, potentially share some functional relationship associated with this perturbation [23, 41, 7, 31].

Here we generated clusters driven by estimating all-against-all Pearson Correlation Coefficients. These clusters serve as sets of putatively co-expressed genes. The dominant biological condition that is potentially driving the clustering is investigated. Putative functional characteristics of these respective sets are inferred by enriched and informative Gene Ontology terms [8, 2]. The method was a collaborative effort and applied in a targetted approach to a collection of publicly available microarray experiments for *Vitis vinifera*. A target list of defence related genes was compiled from the literature based on their respective protein sequences and deduced amino acid sequences [13, 37]. The list of genes are then associated with probe sequences, with the unambiguously associated genes projected onto our co-expressed sets. The methodology allows for the generation of hypothesized functional association between genes, which can then be analysed and investigated further.

3.2 Results and Discussion

We developed the methodology discussed in the Materials and Method section, with biological interpretation performed by Kari du Plessis from the Institute for Wine Biotechnology at Stellenbosch University.

The biological conditions given in the microarray dataset, the respective columns of the expression matrix, were processed and categorized. A particular experiment was annotated and then classified based on information retrieved from Plexdb[8]. The classification involved two major groups and their respective sub-categories. The first group was classified based on treatment which included no treatment, abiotic stress and biotic stress. The second group was classified based on source which consisted of a category for tissue and cultivar specificity and a category for berry development. A summary of the mapping is given in Table 3.1 and Table 3.2.

Several combinations of the model parameters were applied to identify which parameter set provided the optimum level of information to analyse co-expressed sets given the categories above. Putative sets of co-expressed genes where identified by first estimating the all-against-all Pearson Correlation of probesets, then clustering the respective probesets using Markov Clustering. Only clusters, or sets, that contained at least one of our target genes were retained for further analysis. Our target genes were defensin genes, which are genes that have some function associated with plant defense. The categorization of the microarray experiments and the curation of target gene list was

Table 3.1: Summary of sub-categories, based on treatment, used to classify experimental conditions

Category Type	Type of Condition
No Treatment	Biological controls, no treatment controls, berry development with no treatment and samples at 0hr post infection
Biotic Stress	Bios noir infection, Downy mildew infection, leaf roll infection and powdery mildew infection
Abiotic Stress	Water deficit, salinity stress, long or short photoperiod, transgenic over expression and polyethylene glycol treatment

performed by Kari du Plessis from the Institute for Wine Biotechnology at Stellenbosch University.

The clusters indicate subsets of probesets from the original expression matrix. These subsets were then scaled by row and converted to binary values. If given an expression matrix, E, then if a scaled intensity value, E_{ij}, which corresponds to the scaled intensity value of row i and column j, is greater than 0.9 then is was mapped to 1 otherwise it was mapped to 0:

$$\begin{aligned} E^*_{ij} &= 1 \ \textit{if} \ E_{ij} >= 0.9 \\ &= 0 \ \textit{otherwise} \end{aligned}$$

Clusters therefore represented subsets of the binary matrix E^*, say e^*. The category associated with a column of a sub binary expression matrix, was determined to be significant for that cluster if the normalized sum of the column values where greater than some threshold, referred to as a *frequency cutoff*. Several different combinations of cutoff values were investigated and the results presented here are specific combinations determined for the respective experimental category. The specific parameter values were assessed based on the size and information content of the resultant network. The ideal network was a network small enough yet rich enough in content to allow us to hypothesize biological relationships amongst the genes.

Each of the respective subsets were then also analysed by determining their respective informative Gene Ontology terms. These were then used to postulate possible functions associated to our target genes. All the respective networks were visually explored using Cytoscape [29].

Tissue and Cultivar Specific Clusters

We determined which putatively co-expressed set, and therefore target genes, presented evidence for expression specifically associated with plant tissue or

Table 3.2: Experimental Categorization, with description used to classify by treatment and tissue used to classify by source.

Experiment ID	Description	Source
Vv1	Cabernet Sauvignon short term abiotic stress	Shoots, Leaves
Vv2	Long term salt and water stress	Berry
Vv3	Berry differentiation	Berry
Vv5	Chardonnay and Cabernet Sauvignon berry development	Berry
Vv7	Gene Expression of viral diseases in grapevine cultivars	Leaves
Vv9	Temperature effect: Cabernet Sauvignon	Berry
Vv10	Photo period regulation	Buds
Vv11	Pinot Noir berry transcription during ripening	Berry
Vv12	Powdery mildew-induced, transcriptome: Cabernet Sauvignon	Leaves
Vv13	Powdery mildew-induced, transcriptome: Norton	Leaves
Vv15	Berry ripening, gene expression	Berry
Vv16	Berry skin, exogenous abscissic acid	Berry
Vv17	Berry skin, transcriptome of berry cultured *in vitro* treated with abscissic acid	Berry
Vv19	Downy mildew infection, gene expression	Berry
Vv28	Gene expression, viral disease infection	Berry
Vv29	Micro-propagated *Vitis vinifera* transferred to ex vitro conditions	Leaves
Vv31	Expression data, overexpression of VvCBF-4	All aerial tissue

cultivar. Significant expression was determined by a set of parameter values additionally filtering for only unstressed, no treatment, conditions, thus we removed the potential confounding factors associated with expression driven by treatment conditions. This is motivated by [3, 27, 4, 28], where the tissue specificity of defence genes in plants was investigated and it was proposed that these genes are either continuously transcribed or transcription is induced by some external perturbation.

A *correlation cutoff* of 0.8, *inflation value* of 7, *expression cutoff* of 0.9 and *frequency cutoff* of 0.13 was applied in this case and results given in Figure 3.1. From this we determined that seven of our target, defence-related genes, showed evidence of co-expression.

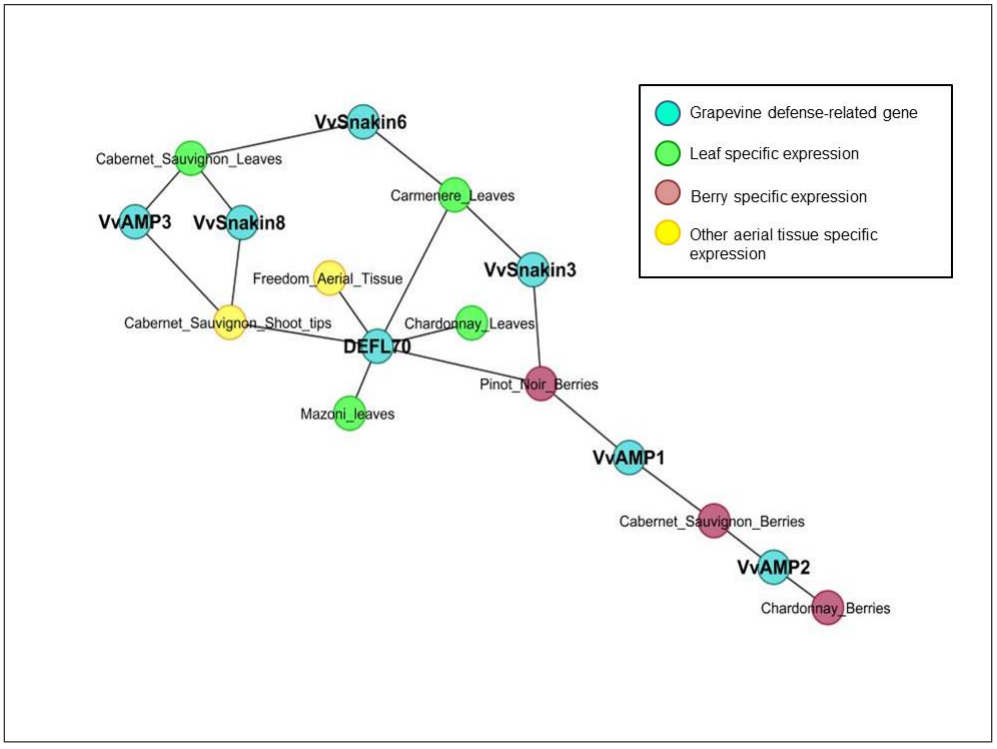


Figure 3.1: Tissue and Cultivar Specific Dominant Conditions

Network indicating the results of applying a parameter set with *correlation cutoff* of 0.8, *inflation value* of 7, *expression cutoff* of 0.9 and *frequency cutoff* of 0.13, then selecting for no-treatment conditions related to tissue and cultivar.

Of these DEFL70 and Vv-Snakin3 showed putative expression in berries and aerial tissues of various cultivars. Here DEFL70 did not present evidence of being putatively expressed for a specific tissue or cultivar. Vv-Snakin8, Vv-AMP3 and Vv-Snakin6 appeared to be putatively expressed in aerial tissues of several cultivars, with Vv-Snakin8 and Vv-AMP3 showing evidence for putative co-expression exclusively for Cabernet Sauvignon. Finally Vv-AMP1 and Vv-AMP2 presented evidence of expression specifically in berries given

no treatment condition in various cultivars. This agreed with evidence found by [9], where a Vv-AMP1 peptide was isolated from a Pinotage cultivar of *Vitis vinifera*. However, this contrasted with findings by [13] that presented evidence for expression of Vv-AMP1 in other tissues for Pinot Noir.

Berry Developmental Stages

Defensin genes, which are known to have function related to the defence of a plant against microbial infections of various stress conditions, also play an important role in the protection of the reproduction potential of a plant, attributed to their abundant presence in seeds and fruits of several plants [35, 24, 20] with berry development and flowering occupying a central position in the grapevine reproductive cycle [6]. Again we analysed for significance based on specifying a parameter set and determining significant conditions associated with berry developmental stages under unstressed conditions and results given in Figure 3.2. Four target genes showed putative co-expression where berry developmental stages were a dominant condition. This was determined with a *correlation cutoff* of 0.7, *inflation value* of 7, *expression cutoff* of 0.9 and a *frequency cutoff* of 0.14. From this Vv-AMP2 presented evidence of putative expression in several stages of berry development. With Vv-AMP1 presenting evidence of expression exclusively in the ripe and ripening stage, showing evidence of putative co-expression with VvSnakin3 during the ripening stage. We also found that DEFL70 was putatively expressed in green and softening stages. The results for both Vv-Snakin3 and Vv-AMP1 are possibly justified by [9, 39], where it was argued that at the start of the ripening stage a certain level of pH is attained, which provides a putatively conductive environment for defence related function.

Abiotic Stress

Abiotic stress can be seen as the negative result that is caused by the influence of non-living perturbations. Stresses can cause the induction of transcription of various genes to respond to the given stress. Defence related genes are known to be potentially induced by signal molecules called hormones, such as abscissic acid. The parameter set used consisted of a correlation cutoff of 0.8, inflation cutoff of 7, expression cutoff of 0.9 and a frequency cutoff of 0.9, see Figure 3.3. From the resulting information we obtained five target genes. Of these five VvAMP1, VvAMP2 and DEFL70 where found to be putatively co-expressed under the stress of water deficiency, with VvAMP1 putatively expressed given abscissic acid. The metabolism of the latter has previously been shown to be connected to water deficiency in [17] and [43]. The interconnection between water stress, salt stress and abscissic acid was also investigated, in several plants, by [34] with specific interest placed on the

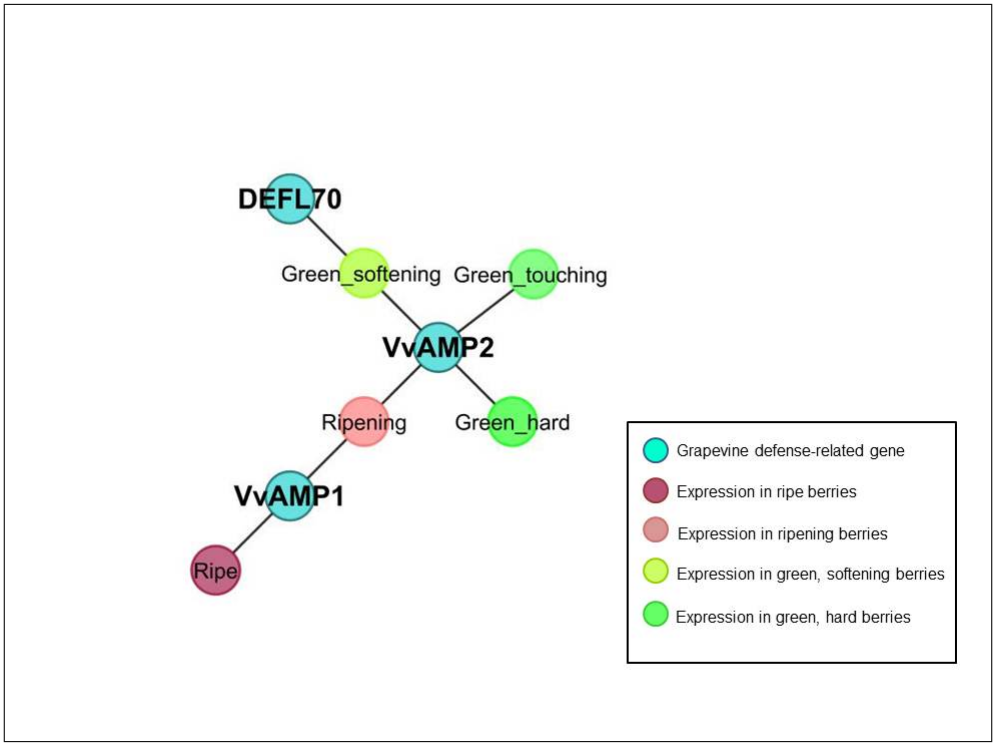


Figure 3.2: Berry Developmental Stages Dominant Conditions

Network indicating the results of applying a parameter set with *correlation cutoff* of 0.7, *inflation value* of 7, *expression cutoff* of 0.9 and *frequency cutoff* of 0.14. Then selecting for no-treatment conditions related to berry development.

effect on plant resistance to pathogens and insects. This may provide further support for the observed putative co-expression.

Biotic Stress

Biotic stress is the negative impact that is induced by the influence of other organisms, including bacteria, fungi and viruses. We applied a *correlation cutoff* of 0.6, an *inflation cutoff* of 7, an *expression cutoff* 0.9 and a *frequency cutoff* of 0.13, resulting in the network given in figure 3.4.

Vv-AMP1 and Vv-Snakin3 showed evidence of putative co-expression given the infection of closterovirus-3. The virus is transmitted by insects and grafting and results in a disease that causes pits and grooves in the trunk, possibly resulting in loss of production [16]. It is hypothesized that these defensin genes induce an inhospitable environment for the insect vector of this virus, evidence for this property of defensin genes is given in [34, 3, 4]. We also found that DELF70 is putatively expressed under the stress of Bois Noir, which causes disease in plants resulting in yellowing of the leaves, slowed growth and low quality fruit [14].

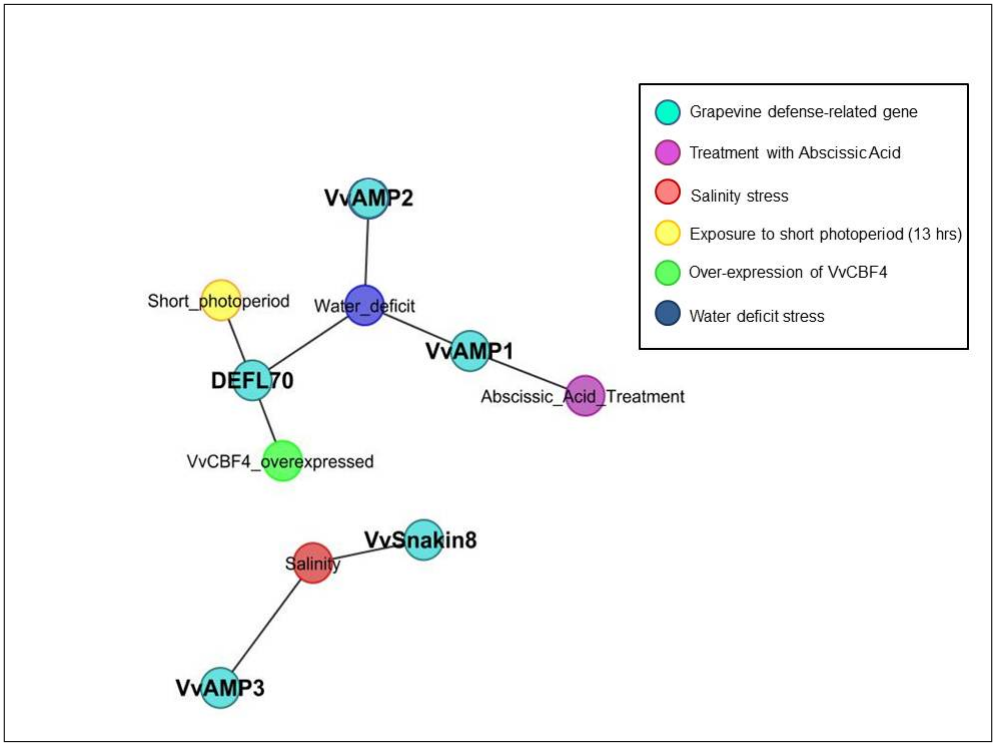


Figure 3.3: Abiotic Stress Dominant Conditions

Network indicating the results of applying a parameter set with correlation cutoff of 0.8, inflation value of 7, expression cutoff of 0.9 and frequency cutoff of 0.9. Then selecting for no-treatment conditions related to abiotic stress.

Gene Ontology

A putative co-expression network was generated using a *correlation cutoff* of 0.8 and *inflation value* of 7. The Markov clusters resulting from these parameters, were mapped to target genes, only those that had at least one target gene were retained for analysis. The Gene Ontology terms associated with these putatively co-expressed genes where obtained using Workbench on the Plaza 2.5 website [25]. The resulting network for Vv-AMP3 was partitioned into three sub-graphs. The networks were annotated, coloured and obtained from [10].

By investigating sub-graphs in this network, we discover three interesting biologically relevant partitions. The first network contains the putative co-expression of Vv-AMP3 with five other genes. Together these genes are involved in various developmental processes, trans-membrane transport activity and external biotic stimuli. Support for the membrane association of Vv-AMP3 is given in [24].

The second network, Figure 3.6, contains Vv-AMP3 with four other putatively co-expressed genes with their underlying GO terms suggesting involvement in nitrogen-related metabolic processes, cation binding, metal ion binding and transcription & translation processes. The ion binding and zinc ion

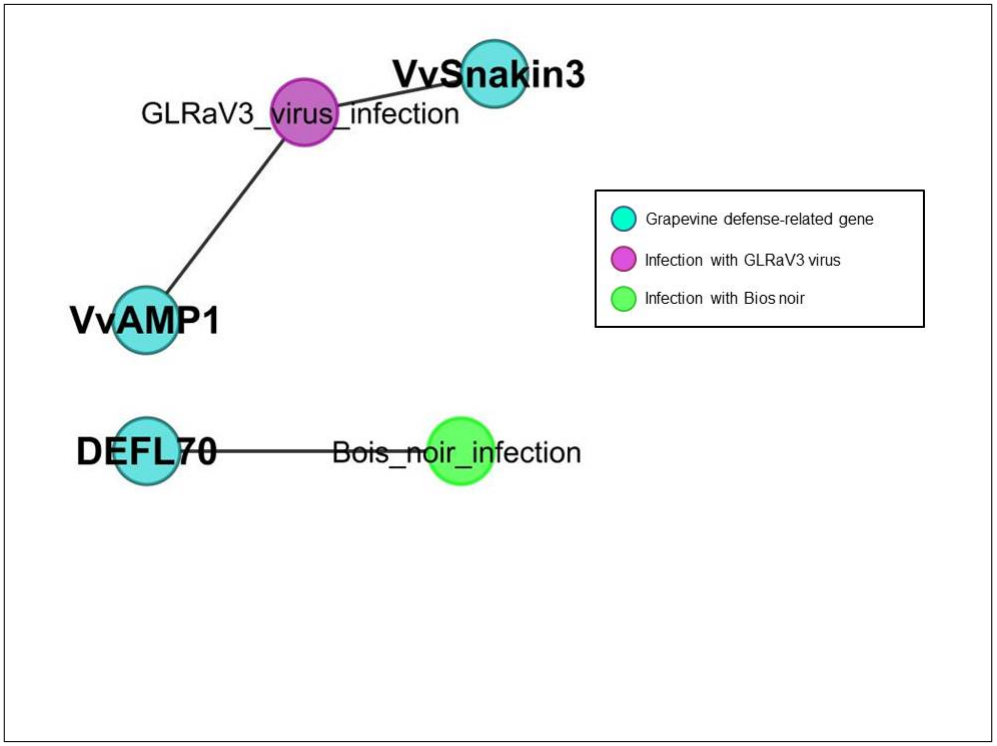


Figure 3.4: Biotic Stress Dominant Conditions

Network indicating the results of applying a parameter set with *correlation cutoff* of 0.6, *inflation value* of 7, *expression cutoff* of 0.9 and *frequency cutoff* of 0.9. Then selecting for no-treatment conditions related to biotic stress.

binding activity of Vv-AMP3 was also discussed in [24] and [21] respectively. The involvement of Vv-AMP3 with regards to nitrogen metabolism was also investigated in [11, 18].

In the third sub-network, Figure 3.7, we have another additional four putatively co-expressed genes involved in protein kinase activity, post-translational modification and nucleotide binding. Evidence for the kinase related activity of Vv-AMP3 was proposed in [26, 22, 38].

3.3 Methods and Materials

Dataset

The raw microarray data files were obtained from the Plexdb website [8], where the data consisted of 18 respective microarray experiments related to *Vitis vinifera*. The files were extracted and normalized with Robust Multiarray Averaging (RMA) using the bioaffy package from Bioconductor in R [15, 12]. This resulted in a expression matrix of $\log_2$ intensity values, where the rows indicated probesets and the columns indicated conditions outlined in Table 3.2. The conditions were then categorized using information obtained from Plexdb

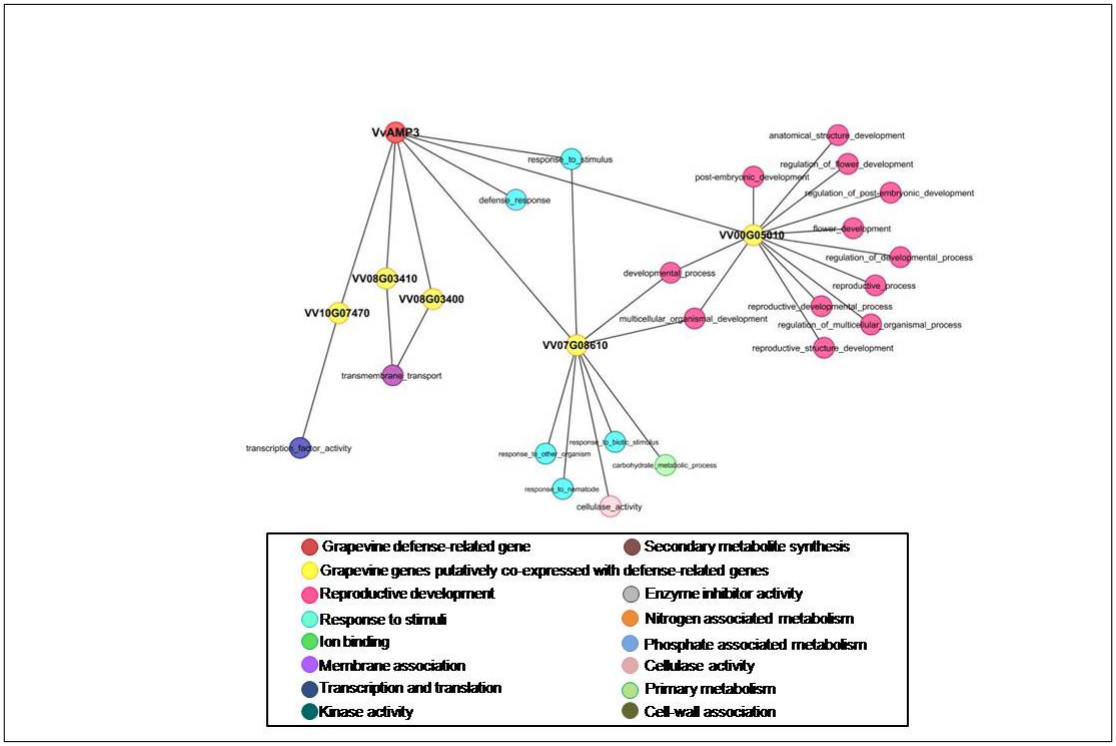


Figure 3.5: First Subnetwork of GO-term annotated Vv-AMP3.

resulting in the categories listed in Table 3.1 and Table 3.2, respectively.

The probesets were then mapped to genes by comparing the respective gene sequences with the probeset sequences using the offline version of NCBI Blast implementation [5]. A probeset was determined to map to a gene if there existed a 100 percent match to the gene for the entire length of the probeset sequence.

Target Gene List

Protein sequences of a total of 97 putatively defence related genes were obtained from [13] and [10]. Pseudogenes, gene fragments and those sequences that were longer than 400bp were removed from the set [10]. The target list of genes was also filtered by a Blast comparison to the probe sequences of the Affy *Vitis vinifera* microarray chip using tBLASTn from Plexdb. A minimum match of 98 percent across the entire length of the probe sequence was used as a cutoff. The resulting genes were compared to the Expressed Sequence Tag database for grapevine, which contain short sequences of cDNA, using a BLASTx analysis. After removing sequences that mapped to the same gene, we obtained our target list.

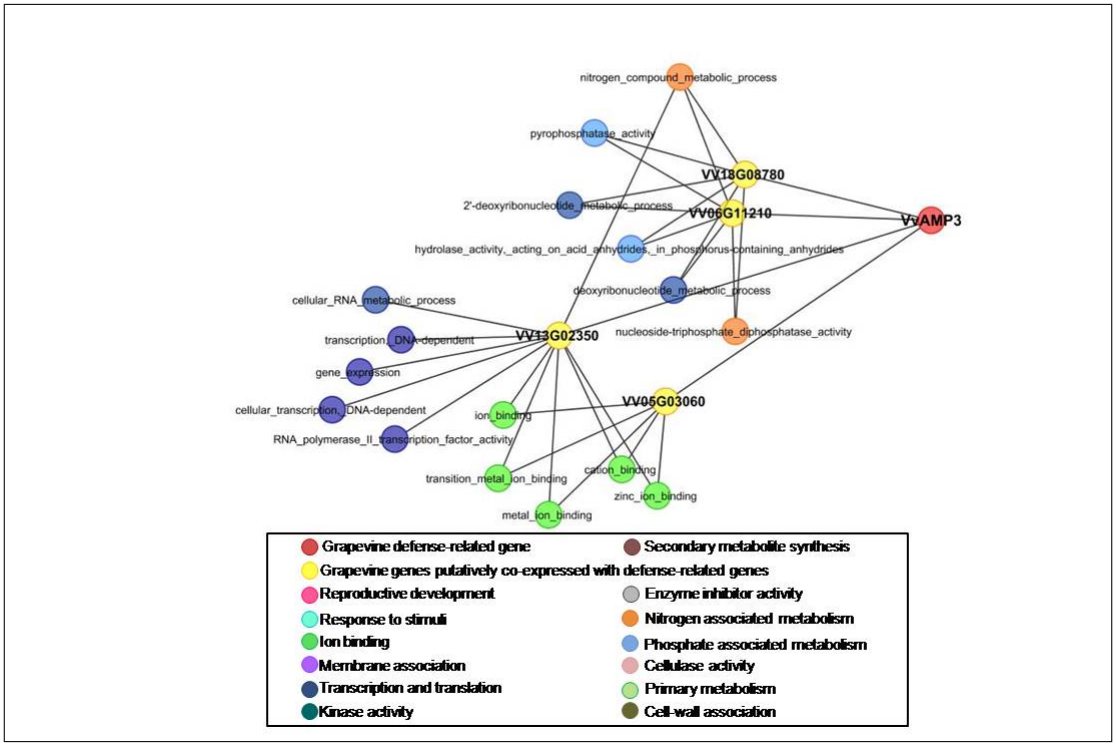


Figure 3.6: Second Subnetwork of GO-term annotated Vv-AMP3.

Pearson Correlation Coefficient

Pearson correlation can potentially capture the similarity between the shapes of two vectors. It has been applied to gene expression data, to infer putative co-expression relationships [19, 42, 30]. It is defined for our purposes as :

Definition 3.1. Pearson Correlation Coefficient:

$$r(\mathbf{x}_j, \mathbf{x}_k) = \frac{\Sigma_{i=1}^{n}(x_{ji} - \bar{x_j})(x_{ki} - \bar{x_k})}{\sqrt{\Sigma_{i=1}^{n}(x_{ji} - \bar{x_j})^2}\sqrt{\Sigma_{i=1}^{n}(x_{ki} - \bar{x_k})^2}}$$

where, $\mathbf{x}_j$ and $\mathbf{x}_k$, are the j^{th} and k^{th} respective variables that are each respective vectors in $\mathbb{R}^n$, $\bar{x_j}$ and $\bar{x_k}$ are the scalars representing the empirical means of the observations of each of the respective j^{th} and k^{th}, variables.

The equation above contained a scaling term, thus when comparing Pearson Correlation Coefficient values we are comparing relative values that are scale invariant. This mitigates the potential bias associated with a vector that has inflated values, such as inflated intensity values in vectors of gene expression resulting from hybridization kinetics. The equation above results in a value between -1 and 1. Here a value of 1 implies perfect correlation, the two vectors have the exact same shape. A value of -1 implies vectors that have the exact opposite shape.

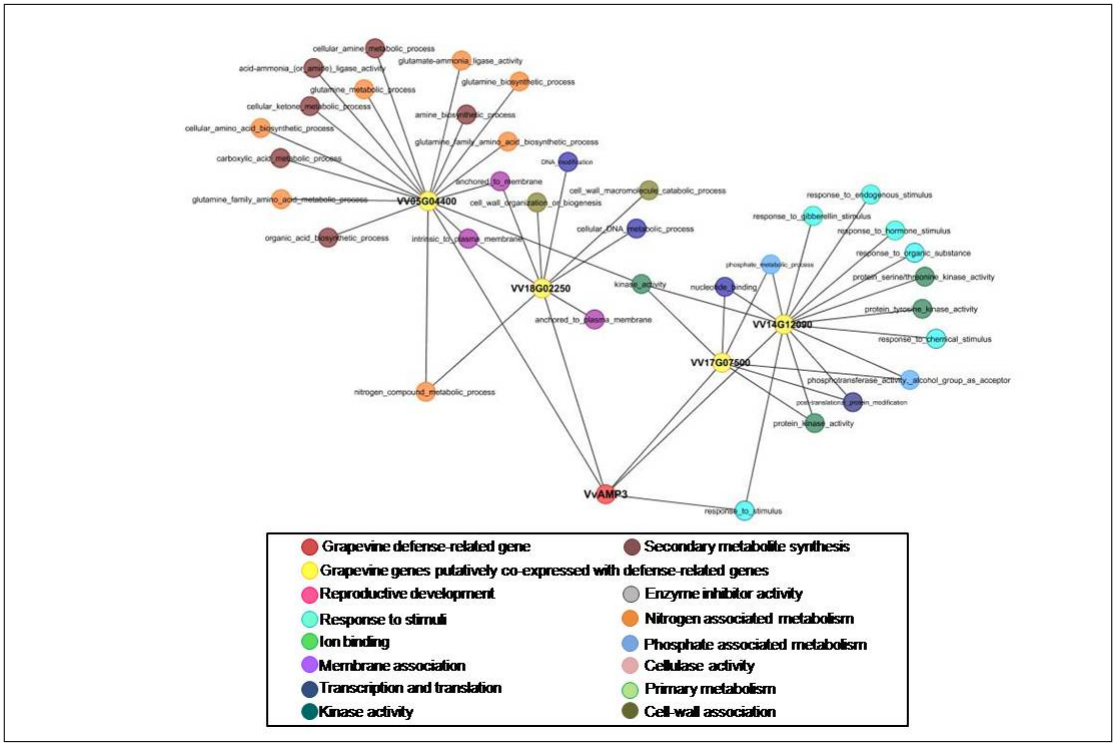


Figure 3.7: Third and Final Subnetwork of GO-term annotated Vv-AMP3.

Two expression sets that have highly similar shapes are said to be putatively co-expressed. In our analysis we performed an all-against-all Pearson Correlation calculation, where our vectors were rows of probeset intensity values given by an expression matrix. To meaningfully analyse the results, we only reported pairs of vectors that have an absolute correlation higher than a particular threshold. This threshold is called a *correlation cut-off*.

Markov Clustering

The stochastic flow that is present in the data can be used to produce clusters, based on a modified random walk. This approach, termed Markov Clustering, was originally introduced in [36]. The process involves two operations that are successively applied in an iterative fashion, an expansion operator and a inflation operator.

Given a similarity matrix, which is a square matrix with elements indicating the similarity between the respective row and column labels. This matrix is transformed into a stochastic matrix by normalizing the columns, such that the sum of the columns are respectively equal to one. An expansion operator is then defined as a random walk of length one, which required the multiplication of a stochastic matrix with itself. The inflation operations involves a parameter called an *inflation value*, which is a scalar greater than 1. Each element of a stochastic matrix is raised, element wise, to the power of this inflation value.

The resulting matrix is then column normalized, resulting in a stochastic matrix. This inflation operation effectively strengthens the stronger stochastic flow between the respective row and column elements, while simultaneously weakening weaker flows.

These two operations are repeated until a level of convergence is achieved, whereby convergence is defined when no significant change is observed after an iteration. This process concludes with a disjoint set of clusters between variables indicated by the respective row and column labels.

Dominant Condition

To determine which are the dominant conditions in a set of putatively co-expressed genes, we first scaled the respective row vectors. These vectors were scaled by the maximum intensity in that vector, thus resulting in a vector with intensity values between 0 and 1. We then mapped the resulting vector to a binary domain, where 1 indicates significant expression under the given condition and 0 otherwise. This was done using a threshold, and scaled intensity higher than a particular threshold (referred to as the *expression cutoff*), was deemed significant.

To extend this to a set, we simply determined the proportion of times that a condition is significant for a given set, which involved the calculation of the column sums of the binary vector, then dividing the results by the number of elements of the set. A condition was then determined to be significant for a particular set if this proportion was greater than a specified threshold, termed the *frequency cutoff*.

3.4 Conclusion and Future Work

We developed a method that putatively captures dominant conditions that may drive the possible co-expression of a set of genes. This method, applied to a microarray dataset for *Vitis vinifera*, was used in a targeted analysis based on putative defensin-like genes.The genes obtained from the literature, and curated using BLAST comparisons, were mapped to a set of unambiguous probesets. These were then analysed using our method in an exploratory analysis. The resulting dominant conditions hypothesized putative contextual interactions for the respective genes, with evidence from the literature supportive of some of these putative interactions.

To further explore and hypothesize biologically relevant functions that may be shared by the sets of putatively co-expressed genes, we analysed the relevant informative Gene Ontology terms of a set. These provided hypothesized functions that may indicate common processes from which gene interactions can be inferred. A literature investigation uncovered support for some of the relevant

informative terms, thus further supporting the gene set structure determined by the method.

The exploratory approach highlighted above can be used to generate hypotheses related to functional annotation and gene interaction. These hypotheses can then be tested to gather evidence for or against the proposed interaction or function. There are several possible improvements that can be made to this methodology. A more rigorous analysis in terms of parameter optimization can be performed. The analysis can also be applied to other eukaryotes for which sufficient information is available. It may be of interest to apply the method in an un-targeted approach to propose a dominant condition driven interaction network from a global context.

3.5 List of References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410.

[2] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J.T. Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, vol. 25, no. 1, pp. 25–29.

[3] Bowles, D.J. (1990). Defense-related proteins in higher plants. *Annual review of biochemistry*, vol. 59, no. 1, pp. 873–907.

[4] Broekaert, W.F., Cammue, B.P., De Bolle, M.F., Thevissen, K., De Samblanx, G.W., Osborn, R.W. and Nielson, K. (1997). Antimicrobial peptides from plants. *Critical Reviews in Plant Sciences*, vol. 16, no. 3, pp. 297–323.

[5] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. (2009). Blast+: architecture and applications. *BMC bioinformatics*, vol. 10, no. 1, p. 421.

[6] Carmona, M.J., Chaïb, J., Martínez-Zapater, J.M. and Thomas, M.R. (2008). A molecular genetic perspective of reproductive development in grapevine. *Journal of experimental botany*, vol. 59, no. 10, pp. 2579–2596.

[7] Curtis, R.K., Orešič, M. and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends in Biotechnology*, vol. 23, no. 8, pp. 429–435.

[8] Dash, S., Van Hemert, J., Hong, L., Wise, R.P. and Dickerson, J.A. (2012). Plexdb: gene expression resources for plants and plant pathogens. *Nucleic acids research*, vol. 40, no. D1, pp. D1194–D1201.

[9] De Beer, A. and Vivier, M.A. (2008). Vv-amp1, a ripening induced peptide from vitis vinifera shows strong antifungal activity. *BMC plant biology*, vol. 8, no. 1, p. 75.

[10] Du Plessis, K. (2012). Personal Communication.

[11] Espinoza, C., Medina, C., Somerville, S. and Arce-Johnson, P. (2007). Senescence-associated genes induced during compatible viral interactions with grapevine and arabidopsis. *Journal of Experimental Botany*, vol. 58, no. 12, pp. 3197–3212.

[12] Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004). affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, vol. 20, no. 3, pp. 307–315. ISSN 1367-4803.

[13] Giacomelli, L., Nanni, V., Lenzi, L., Zhuang, J., Serra, M.D., Banfield, M.J., Town, C.D., Silverstein, K.A., Baraldi, E. and Moser, C. (2012). Identification and characterization of the defensin-like gene family of grapevine. *Molecular plant-microbe interactions*, vol. 25, no. 8, pp. 1118–1131.

[14] Hren, M., Nikolić, P., Rotter, A., Blejec, A., Terrier, N., Ravnikar, M., Dermastia, M. and Gruden, K. (2009). 'bois noir'phytoplasma induces significant reprogramming of the leaf transcriptome in the field grown grapevine. *BMC genomics*, vol. 10, no. 1, p. 460.

[15] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, vol. 4, no. 2, pp. 249–264.

[16] Kotze, A.C. (2007). *Anomalous PCR results to Grapevine leafroll associated closterovirus type 3 in South Africa*. Ph.D. thesis, University of Pretoria.

[17] Koyama, K., Sadamatsu, K. and Goto-Yamamoto, N. (2010). Abscissic acid stimulated ripening and gene expression in berry skins of the cabernet sauvignon grape. *Functional & integrative genomics*, vol. 10, no. 3, pp. 367–381.

[18] Lam, H.-M., Coschigano, K., Oliveira, I., Melo-Oliveira, R. and Coruzzi, G. (1996). The molecular-genetics of nitrogen assimilation into amino acids in higher plants. *Annual review of plant biology*, vol. 47, no. 1, pp. 569–593.

[19] Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, L., Chen, L., Chen, T.M., Chin, M.C., Chong, J., Crook, B.E., Czaplinska, A., Varnam, L.R., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T.A., Donelan, M.J., Dong, H.-W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B.A., Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R., Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R.A., Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramee, A.R., Larsen, K.D., Lau, C., Lemon, T.A., Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng, L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S.A., Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N.V., Sivisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K.A., Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpf, K.-R., Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., Wood, M., Yaylaoglu, M.B., Young, R.C., Youngstrom, B.L., Yuan, X.F., Zhang, B., Zwingman, T.A. and Jones, A.R. (2006). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, vol. 445, no. 7124, pp. 168–176.

[20] Meyer, B., Houlne, G., Pozueta-Romero, J., Schantz, M.-L. and Schantz, R. (1996). Fruit-specific expression of a defensin-type gene family in bell pepper (upregulation during ripening and upon wounding). *Plant physiology*, vol. 112, no. 2, pp. 615–622.

[21] Mirouze, M., Sels, J., Richard, O., Czernic, P., Loubet, S., Jacquier, A., François, I.E., Cammue, B., Lebrun, M., Berthomieu, P. and Marques, L. (2006). A putative novel role for plant defensins: a defensin from the zinc hyper-accumulating plant, arabidopsis halleri, confers zinc tolerance. *The plant journal*, vol. 47, no. 3, pp. 329–342.

[22] Miya, A., Albert, P., Shinya, T., Desaki, Y., Ichimura, K., Shirasu, K., Narusaka, Y., Kawakami, N., Kaku, H. and Shibuya, N. (2007). Cerk1, a lysm receptor kinase, is essential for chitin elicitor signaling in arabidopsis. *Proceedings of the National Academy of Sciences*, vol. 104, no. 49, pp. 19613–19618.

[23] Mootha, V., Lindgren, C., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E. Ridderstrale, M. and Laurila, E. (2003). Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, vol. 34, pp. 267–273.

[24] Osborn, R.W., De Samblanx, G.W., Thevissen, K., Goderis, I., Torrekens, S., Van Leuven, F., Attenborough, S., Rees, S.B. and Broekaert, W.F. (1995). Isolation and characterisation of plant defensins from seeds of asteraceae, fabaceae, hippocastanaceae and saxifragaceae. *FEBS letters*, vol. 368, no. 2, pp. 257–262.

[25] Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y. and Vandepoele, K. (2009). Plaza: a comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell Online*, vol. 21, no. 12, pp. 3718–3731.

[26] Robatzek, S., Chinchilla, D. and Boller, T. (2006). Ligand-induced endocytosis of the pattern recognition receptor fls2 in arabidopsis. *Genes & Development*, vol. 20, no. 5, pp. 537–542.

[27] Ryan, C.A. and Jagendorf, A. (1995). Self defense by plants. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 10, p. 4075.

[28] Ryan, C.A. and Moura, D.S. (2002). Systemic wound signaling in plants: a new perception. *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6519–6520.

[29] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, vol. 13, no. 11, pp. 2498–2504.

[30] Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, vol. 302, no. 5643, pp. 249–255.

[31] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550.

[32] Team, B.C. (2002). Bioconductor-open source software for bioinformatics.

[33] Team, R.D.C. (2005). R: A language and environment for statistical computing.

[34] Thaler, J.S. and Bostock, R.M. (2004). Interactions between abscisic-acid-mediated responses and plant resistance to pathogens and insects. *Ecology*, vol. 85, no. 1, pp. 48–58.

[35] Thomma, B.P., Cammue, B.P. and Thevissen, K. (2002). Plant defensins. *Planta*, vol. 216, no. 2, pp. 193–202.

[36] van Dongen, S.M. (2000). Graph clustering by flow simulation.

[37] Vivier, M.A. and Pretorius, I.S. (2002). Genetically tailored grapevines for the wine industry. *TRENDS in Biotechnology*, vol. 20, no. 11, pp. 472–478.

[38] Wan, J., Zhang, X.-C., Neece, D., Ramonell, K.M., Clough, S., Kim, S.-y., Stacey, M.G. and Stacey, G. (2008). A lysm receptor-like kinase plays a critical role in chitin signaling and fungal resistance in arabidopsis. *The Plant Cell Online*, vol. 20, no. 2, pp. 471–481.

[39] Wang, S., Rao, P. and Ye, X. (2009). Isolation and biochemical characterization of a novel leguminous defense peptide with antifungal and antiproliferative potency. *Applied microbiology and biotechnology*, vol. 82, no. 1, pp. 79–86.

[40] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Genes: Structure, replication and expression. In: *Prescott's microbiology*, chap. 12. McGraw-Hill.

[41] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Microbial genomics. In: *Prescott's microbiology*, chap. 16. McGraw-Hill.

[42] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, p. 1128.

[43] Zhu, J., Verslues, P.E., Zheng, X., Lee, B.-h., Zhan, X., Manabe, Y., Sokolchik, I., Zhu, Y., Dong, C.-H., Zhu, J.-K., Hasegawa, P.M. and Bressan, R.A. (2005). Hos10 encodes an r2r3-type myb transcription factor essential for cold acclimatization in plants. vol. 102, no. 28, pp. 9966–9971.

Chapter 4

Cross Cluster Gene Interaction Detection

4.1 Introduction

Understanding how genes interact with each other in a biological context is critical to understanding how a cell develops and functions. Genes generally code for proteins, which are translated from mRNA, and these proteins are the basic building blocks that are responsible for cellular structure and function. Large scale experiments such as microarrays allow us to view a snapshot of the mRNA activity under a particular set of perturbations. This activity, in some cases, can be viewed in the context of whole genomes [49].

These types of experiments have contributed significantly to our current understanding of biological phenomena [24, 20]. The datasets, however, are often very noisy and classical data mining techniques often lack sufficient power to provide meaningful information. The problem stems from the large number of variables involved in an experiment, involving often a low number of conditions, in a system in which the interaction of variables may be very complex. This is often compounded by a low number of repeated samples [35, 3]. Delineating the interactions that genes have in a biological context from these data sources may improve our understanding of these complex interactions, and therefore allow for improved understanding of these systems.

Patterns in such large datasets are often studied by generating modules of related variables by using clustering. This concept has been applied to microarray data to generate sets of putatively co-expressed, or co-regulated, genes [2, 14, 10]. These modules can then be used to understand the relation between different variables, or genes, in the system.

It is also convenient to view this hypothesized complex combinatorial set relation in the context of a network. In the network, $G = (V, E)$, genes are the node set $V \in G$, and the problem becomes inferring the edge structure $E = \{v_i, v_j\}; v_i, v_j \in V$. These networks, often termed gene interaction networks or

gene regulatory networks, have received a lot of attention recently, a review of some of these methods and their respective application can be found in [13, 22, 4]. These approaches attempt to infer the structure of these networks using a variety of methodologies.

Some approaches attempt to model the dynamics of gene interactions using models of chemical kinetics, such as a system of ordinary differential equations [41, 12]. These models often required a large number of parameters to be estimated and the structure of the equations may require prior knowledge of biological complexity.

There is also a particular class of models, called Bayesian Networks, that model the joint probability distribution of the biological phenomena, where the nodes are viewed as random variables and the data is an observation of this joint distribution. This joint probability distribution can be factored into conditional probability distributions based on the dependence structure outlined by the network structure, a factorization based on statistical dependence. This structure is often inferred by casting the problem in the context of a Bayesian inference model and sampling probable graph structure from a posterior distribution [18]. Several of these proposed models are not necessarily distinct from one another as there is evidence to suggest that there is a relationship between a system of ordinary differential equations and a Dynamic Bayesian Network[33].

Here we illustrate an exploratory method that proposes a network structure for a gene interaction network. We use Markov Blankets to determine the statistical dependence structure, in a network context, between modules of putatively co-expressed genes, clustered using Markov Clustering. We apply this method, based on the model proposed by [29], to time-series gene expression data obtained from real microarray experiments, exploring the inferred structure using real world knowledge of biologically relevant interactions. Networks, or graphs, are visualized using Cytoscape 2.8 [44]

4.2 Results and Discussions

To infer the edge structure of gene-gene interaction networks the conditional statistical dependence for each probeset was modelled by inferring their respective Markov Blankets, using an approach highlighted by Li et al [29]. Candidates for inclusion in a probeset's Markov Blanket were selected from disjoint sets of clustered putatively co-expressed probesets. These disjoint sets were calculated using Pearson Correlation-based Markov Clustering (see Materials and Methods). The underlying hypothesis was that a subset of these putatively co-expressed probesets, or genes, are the child nodes of some subset of parent nodes from another set of putatively co-expressed probesets. Thus we hypothesized that biologically relevant gene-gene interactions can be inferred under the assumption that putatively co-expressed genes have some common

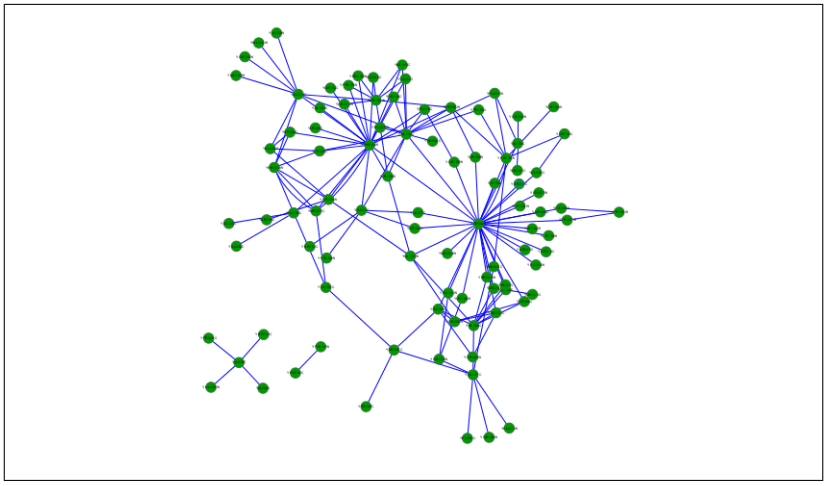


Figure 4.1: Cross Cluster Gene Interaction Network

Nodes indicate genes and edges connect nodes that are significantly statistically dependent across clusters

signal that can be measured by statistical dependence. It is therefore these signals that we attempted to infer. We used the implementation in *bnlearn* package in R of the Markov Blanket detection algorithm [42, 46].

In order to explore the hypothesis underlying our methodology, or gather evidence to support it, we applied our method to a publicly available microarray dataset obtained from NCBI [15]. The dataset is a *Saccharomyces cerevisiae* time course gene expression dataset concerned with the ageing of non-dividing yeast cells without the application of caloric restrictions (see Materials and Methods) [15]. *S. cerevisiae* is a well-studied organism and we therefore benefit from a wealth of validated information that can be used to improve the confidence in our approach.

After normalizing the dataset using Robust Multichip Averaging (RMA) [26] and obtaining the log_2 expression matrix of probesets, we calculated the all-against-all row-wise Pearson Correlation Coefficients. We then selected a subset of probesets for further analysis, based on genes that matched those in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) metabolic pathway for *Saccharomyces cerevisiae* [34]. Only row pairs that had an absolute correlation of 0.9 and higher were selected and then clustered using Markov Clustering with an inflation value of 7 [47]. Then for each element in each cluster the Markov Blanket for that element was estimated from elements of every other cluster, respectively. For the Markov Blankets a 0.05 level of significance was applied, with corrections for multiple hypothesis testing performed using the Benjamin-Hochberg procedure to control the false discovery rate [5]. The resultant network is depicted in Figure 4.1

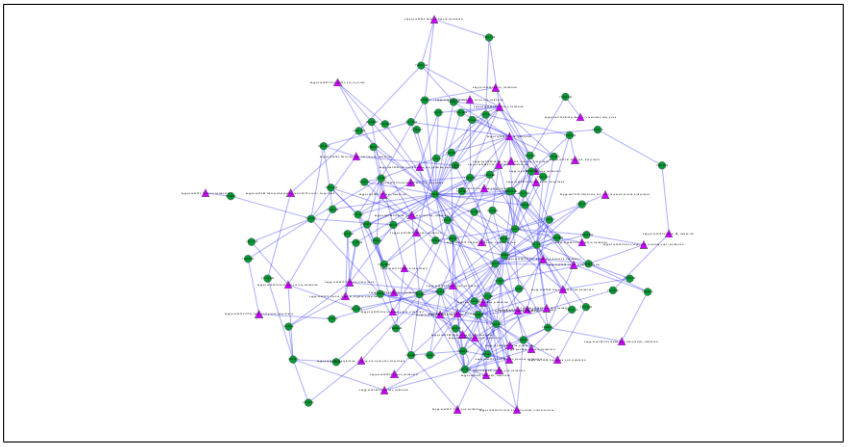


Figure 4.2: Annotated Cross Cluster Gene Interaction Network

Nodes are KEGG metabolic pathways (purple triangles) and genes (green circles). Edges between genes are the result of the Markov Blanket analysis and an edge between genes and a pathway indicates genes occurring in that pathway

4.2.1 Projection on Metabolic Pathways

The directionality was removed from the network inferred from our methodology, thus resulting in undirected edges between particular gene pairs. We annotated the network with their respective KEGG metabolic pathways by adding nodes indicating the respective pathway and then joining these pathways to a gene if the particular gene is involved in the pathway [34](see Materials and Methods). The network obtained from the KEGG metabolic pathways is given in Figure 4.2

In order to identify meaningful edges in our originally inferred network, we identified cliques of size 3 or larger in the original network after it was annotated effectively attempting to identify triangular structures. These cliques indicated that an edge inferred by our network was between two genes in the same pathway. The Fisher Exact test was then used to determine if the number of cliques found in this annotated network was statistically significant (see Materials and Methods).

We found a total of 17 cliques that were of size 3 or larger, with the corresponding p-value from the Fisher Exact test of 0.026 which was determined significant at a 0.05 significance level, the cliques along with their respective elements are given in Figure 4.3. There was sufficient evidence to reject the null-hypothesis of independence at a 0.05 significance level. The cliques indicated that their edges captured putatively potential pathway effects estimated from the data.

In the topology of the clique network structure there were several cliques that are neighbours of the *glycolysis/gluconeogenesis* pathway, one of the major production sources for energy in yeast metabolism [49]. The neighbouring pathways include *fructose and mannose metabolism*, *galactose metabolism* and

pyruvate metabolism. *Galactose* is a known input to *glycolysis* and *pyruvate* a known output [48]. The effect of *gluconeogenesis* on the chronological ageing of yeast is investigated in [31]. This investigation suggested that a shift from *glycolysis* to *gluconeogenesis* is associated with the ageing of yeast cells. Furthermore, this ageing also involved an increase in activity of *fructose 1,6 biphosphatase* which has been shown to be negatively affected by YMR205C [23, 31].

The clique structure given by Clique 14 is associated with *glycine and syrine metabolism* and is also a neighbour of the *glycolysis/gluconeogenesis* pathway. The genes associated with the former pathway were determined to be putatively related to the starvation phenotype and may play an important role in the survival of yeast cells [38]. It was also determined in the same study that genes associated with *glycolysis/gluconeogenesis* may have a similar putative relationship with the starvation phenotype. Furthermore, the significant clique-based pathways suggested that the yeast cells may be metabolically active. This agreed with the findings of [15] (unpublished), where from initial expression analysis it was determined that the *glycolytic* genes were up-regulated along with their trans-acting regulators.

The involvement of the gene YHR104W, also known as GRE3, in the lifespan of yeast cells was predicted and putatively validated in [52]. It was proposed that the influence of GRE3 on the *pentose phosphate* pathway effects reactive molecules that contain oxygen (ROS) and that ROS has a putative association with the chronological ageing of yeast cells [40, 52].

In the network YOR374W, also known as *Aldehyde Dehydrogenase* (ALD4), has a high degree. This gene is involved in several pathways and plays an important role in ethanol and pyruvate metabolism amongst others [7]. A yeast deletion strain for this particular gene resulted in an increased survival during ageing when compared to the wild-type, suggesting that ALD4 may play a putative role in the cellular ageing [27].

4.3 Materials and Methods

4.3.1 Data

The microarray dataset used to evaluate the performance of our approach was obtained from the publicly available NCBI GEO database [15]. The experiment was performed by Nagarajan et al (unpublished) [15], and sought to investigate the chronological ageing of yeast without enforcing caloric restrictions. The chronological age of cells is, in general, defined as the lifespan of non-dividing cells. There have been several studies done to investigate the relation between chronological ageing and the application of caloric restrictions. From these studies it has been established that the longevity of cells improves when caloric restrictions are applied [30, 32]. Delineating the mechanisms in-

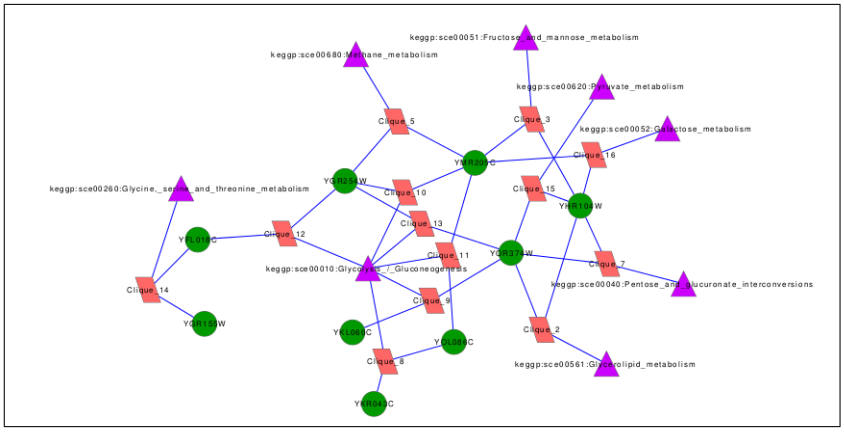


Figure 4.3: Cliques Graph

Cliques of size 3 or larger centred around KEGG pathways estimated from the Annotated Cross Cluster Gene Interaction Network. Nodes indicate KEGG pathways (purple triangles), genes (green circles) and clique identifiers (pink parallelograms). Edges indicate a relation of genes and pathways with particular cliques, respectively

volved in the observed increase in longevity is a continuing area of research. In the experiment underlying the microarray dataset, the authors sought to study the effect that subjecting these non-dividing yeast cells to excess nutrition would have. The cells were immobilized by encasing them in alginate and then they were fed nutrients in a temperature controlled bioreactor. This study was performed over 17 days, with sampling done in triplicate on day 1, 3, 5, 10 and 17, respectively. RNA sampling for the microarray experiment was also done on the pre-immobilized state. This included the batch growth phase, chemostat growth phase and the day prior to immobilization. The raw files were obtained from NCBI and normalized using Robust Multichip Averaging [26]. We also performed a set of preliminary analyses to assess the quality of the data; these were all done using the Bioconductor package in R [19, 45, 46].

The expression matrix obtained after normalization was then averaged over the repeats of the samples retrieved during the immobilization state. We only concerned ourselves with the time series over these immobilization states, thus we obtained a matrix where the rows indicated probesets and the columns indicated the 5 respective time steps.

There is some ambiguity that needs to be considered with regards to the mapping of probesets to their respective genes. To keep our analysis as unambiguous as possible we performed our analysis primarily on probesets. We only mapped the probesets to genes prior to validation. The mapping was performed by using a Blast comparison of the probe sequences, used as our query, to the respective gene sequences. The Blast comparison was done using the offline version of NCBI blast [1, 9]. If there existed a 100% match between probe and gene over the entire length of the query, then we concluded that the probe mapped to the particular gene. This resulted in the many-to-many map in Figure 4.4

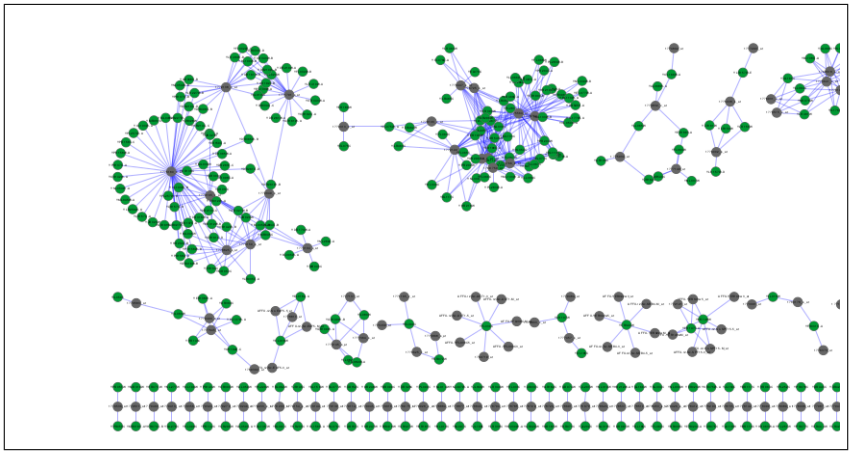


Figure 4.4: Probeset Ambiguity

Partial image of a graph indicating probe to gene mapping ambiguity. Nodes are probesets (grey circles) and genes (green circles), respectively. Edges indicate similarity between a gene and a probeset.

4.3.2 Pearson Correlation Coefficient

Each probeset, indicated by a row in the normalized expression matrix, can be viewed as a vector. With a vector the Pearson Correlation Coefficient [37] can be used as a metric to determine the similarities between the expression patterns of these probesets, with the Pearson Correlation Coefficient defined as:

Definition 4.1. Pearson Correlation Coefficient:

$$r(\mathbf{x}_j, \mathbf{x}_k) = \frac{\Sigma_{i=1}^{n}(x_{ji} - \bar{x_j})(x_{ki} - \bar{x_k})}{\sqrt{\Sigma_{i=1}^{n}(x_{ji} - \bar{x_j})^2}\sqrt{\Sigma_{i=1}^{n}(x_{ki} - \bar{x_k})^2}}$$

where, $\mathbf{x}_j$ and $\mathbf{x}_k$, are the j^{th} and k^{th} respective variables that are each respective vectors in $\mathbb{R}^n$, $\bar{x_j}$ and $\bar{x_k}$ are the scalars representing the empirical means of the observations of each of the respective j^{th} and k^{th}, variables.

This metric measures the similarity between the directions of two vectors by returning a value between -1 and 1. Here a value of 1 (-1) indicates perfect (imperfect) correlation. Thus, expression patterns that are highly similar will have a Pearson Correlation Coefficient closer to 1. The normalization in the calculation of the Pearson Correlation Coefficient results in the metric being invariant to scale. This is an attractive property of the metric as it may mitigate the possibility of modelling artefacts such as hybridization kinetics. For our analysis we defined expression patterns as similar by using the absolute value of the Pearson Correlation Coefficient, thus we removed the distinction between similar and inversely similar expression patterns.

Correlation as a metric has been used before in the analysis of gene expression data [25, 39]. The measure intuitively captures the idea of similar

expression profiles and has therefore also been utilized as a distance metric for clustering of gene expression profiles [51, 16, 28].

4.3.3 Markov Clustering

Clustering can be used as a means to uncover the patterns present in the data. By generating meaningful groups of variables we can potentially obtain modules of co-expression. These modules will allow us to putatively infer biologically meaningful interactions, thus inferring the edge structure of a gene-gene interaction network.

We used the Markov Clustering algorithm of [47], to cluster the respective probesets into sets based on an all-against-all absolute Pearson Correlation Coefficient matrix, where this correlation matrix was utilized as a similarity matrix.

The algorithm produces clusters by modelling the statistical flow within the dataset based on a stochastic matrix obtained from some pre-calculated similarity matrix. A stochastic matrix is the result of performing column-wise normalization on a matrix, such that each column sums to one. In the algorithm two primary operations are used iteratively throughout the algorithm until an acceptable level of convergence is obtained. One operation simulates a random walk, using matrix multiplication of a stochastic matrix with itself. The other operation amplifies stronger flows and represses weaker ones by using element-wise exponentiation, where the exponent used is a parameter called the inflation value. The latter operation concludes by transforming the resultant matrix into a stochastic matrix. The algorithm terminates when no significant change in the matrix obtained is observed from the application of either respective operations. The end result of applying the algorithm to a correlation matrix of probeset expression values is a set of disjoint clusters, with each cluster representing putatively co-expressed probesets.

4.3.4 Markov Blankets

In general, a Markov Blanket describes a minimal set of statistical conditional dependences. In the context of network structural inference these define a neighbourhood for a particular node, when the network is viewed in terms of a joint probability distribution. More formally given a set of variables or nodes, say V, the Markov Blanket for a variable $x \in V$, denoted $MB(x)$, is defined as a minimal set such that:

$$\forall y \in V, y \in MB(x) \textit{ iff } x \textit{ is conditionally dependent on } y \textit{ given } V/\{x, y\} \tag{4.3.1}$$

Intuitively the above defines a cloud of information required to infer the probability distribution of a particular variable, a cloud that blocks the flow of information between the variable and the rest of network. This concept

has been used in the problem of inferring the structure of networks, such as gene regulatory networks. In particular, [29], described an approach that combines the application of a set of ordinary differential equations with a Markov Blanket discovery algorithm to infer the network structure of a gene regulatory network. From the work of of [29] and [6], gene expression here is assumed to follow a system of autoregressive equations as follows:

$$X(t+1) = B * X(t) \tag{4.3.2}$$

Here, $X(t)$, is a vector of log_2 expression levels of genes at time point t. The effect that gene j has on gene i is measured by the ij^{th}element of the square coefficient matrix B.

We infer the Markov Blanket across time; therefore we capture information that is propagated forward in time, which is justified by potential gene interactions occurring over time. This effectively means that we are modelling a Markov Process, were our goal is to estimate the dependency structure of this process. We determine $MB(X_i(t+1))$, where the blanket itself will consist of a subset of $X(t)$.

Let X be an expression matrix of probesets where columns indicate time, such that $X(t)$,$t \in 1 \dots T$, is the t^{th} column which is a vector of log_2 expression values for the probesets at time t. Let $X^{(c)}$, be the expression matrix consisting only of the rows corresponding to probesets in cluster c, with $X^{(c)}_{-t}$, as the aforementioned expression matrix with the column corresponding to time point t, removed. We then estimate the Markov Blanket for the variables using Algorithm 1, which is a modified version of the algorithm proposed by [29]. In [29], the Markov Blanket for a particular target variable was inferred by using all other variables as potential parents. In our case, we partition this set of all other variables not in the same cluster as the target variable into disjoint sets, exploring each set individually as a set of potential parents. Effectively we partition the search space, generating a smaller background against which to determine the parents for a target variable. The Markov Blanket itself is determined using the Fast Incremental Association Markov Blanket (fast-IAMB) algorithm of [50], with a shrinkage estimator for mutual information as a conditional independence test. Mutual information is a measure of the mutual dependence between two stochastic variables and can be defined in terms of entropy, see [36]. Here entropy refers to measure of uncertainty related to a stochastic variable originally introduced by Shannon [43]. The James-Stein shrinkage estimate for mutual information used in this case is a regularized estimate of mutual information based on estimating entropy, for more information on this estimate see [21].

```
Algorithm 1:
for child in set_of_variables:
  child_vector = [X_i(2),... X_i(T)]
  MB(child) = ∅
  for c in all_the_clusters such that child not in c:
    potential_parents = X_{-T}^{(c)}
    MB^(c)(child)= estimated Markov Blanket for child
                           from potential_parents
                           by comparing vectors.
    MB(child) = MB(child) ∪ MB^(c)(child)
return all Markov Blankets
```

4.3.5 Cliques

A clique in terms of an undirected graph, is a complete subgraph. Therefore a clique is a collection of nodes, such that there is an edge between every node pair. A maximal clique is a clique that cannot be extended by adding another node. We used a version of the algorithm originally proposed by [8], discussed in detail in [11], to find all the maximal cliques in our undirected graph. We only determined cliques that are centred around KEGG pathway nodes, in the annotated version of our inferred network. We also considered cliques that were of size 3 or larger, as this minimal triangular shape indicates an edge, or edges, we inferred that may have putative meaning in a metabolic context.

4.3.6 Kyoto Encyclopaedia of Genes and Genomes

Files containing information on the pathways and their respective genes were downloaded in xml format from KEGG on June 27, 2011 [34]. These files were parsed using a Perl script, resulting in a network given in Figure 4.5. Nodes in the network are genes and pathways, with the respective edges connecting a gene and pathway if and only if the gene is contained in that pathway.

4.3.7 Fisher Exact Test

The Fisher Exact Test [17], can be used to determine if there is an association between the column and row classifications of a contingency table. An example of a general contingency table is given in Table 4.1:

Under the assumption of independence, the contingency table indicated in Table 4.1 follows a Hypergeometric distribution, given by:

$$\frac{\binom{d}{x}\binom{n-d}{a-x}}{\binom{n}{a}} \tag{4.3.3}$$

The Fisher Exact Test determines, under the null-hypothesis of independence, the number of contingency tables that can be generated, using the same

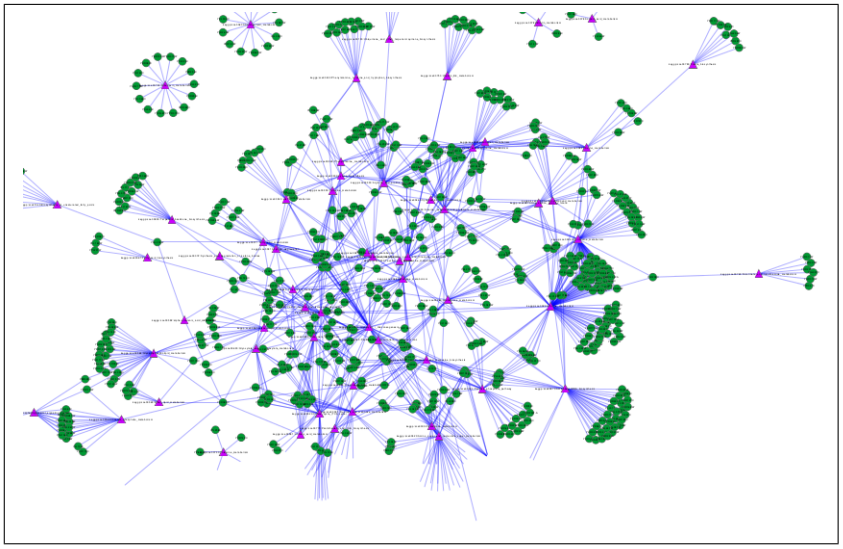


Figure 4.5: KEGG Metabolic Pathway Network

Nodes indicate KEGG metabolic pathways (purple triangles) and genes (green circles). Edges between genes and a pathway indicate genes occurring in that pathway

	Element in Category A	**Element in complement of Category A**	
Element in Category B	x	d - x	d
Element in Complement of Category B	a -x	n+x-a-d	n-d
	a	n-a	n

Table 4.1: An example of a contingency table. Elements in the table are the number of variables that have occurred in both the corresponding row and column labels.

marginal distributions, with equally or more extreme values than the observed. The null hypothesis, in the context of Table 4.1, is that there is no difference in the proportion of events of category A that also fall in B and the proportion of events of category A that do not fall in B. In our case we applied an alternative hypothesis that is two-sided. Rejecting the null hypothesis therefore suggested that there is significant evidence to reject the notion of no association between the categories A and B.

With regards to the clique comparison, category A was replaced by Edges in Inferred Network with its complement as Edges not in Inferred Network. Category B was replaced by Edges mapped to a Pathway Clique and its complement Edges not mapped to a Pathway Clique.

4.4 Conclusion and Future Work

Our method appears to capture putative pathway-based interactions between genes, when given a metabolic background. We hypothesized that clusters of putatively co-expressed genes may have some common upstream interaction. Under this hypothesis we generated sets of putatively co-expressed genes and modelled the element-wise statistical dependence between the respective sets, over time. We also proposed an approach that can be used to putatively explore the context of the inferred edge structure and gain support from known biological interactions.

Our method was specifically applied to a microarray experiment involving over-fed immobilized yeast cells. Performing a contextual analysis on the inferred topology, in the context of yeast metabolism, we discovered evidence suggesting that the yeast cells may be active in terms of energy metabolism. This agreed with the findings of the original curators of the dataset. We have also identified components that may play a role in the cellular ageing of yeast cells. The connections between these components, which are given by the inferred edge structure, may therefore propose novel relationships between these components that can be more rigorously investigated.

In partially exploring the topological structure of the inferred network, we have found that these structures tend to support the interaction of genes in well established pathways. This tends to support the validity of the approach and suggests that there may be several novel interactions hypothesized by the entire network. This may become more evident with further exploration of the network structure or alternatively a genome wide inferred network. There are also alternative ways to explore the topology of the network, other than the clique approach used above, that may highlight or propose novel hypotheses that are biologically meaningful.

4.5 List of References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410.

[2] Bar-Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503.

[3] Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K. and Jaakkola, T.S. (2003). Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10146–10151.

[4] Barabási, A.-L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113.

[5] Benjamin, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Statistical Society, Series B (57)*, vol. 289.

[6] Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S. and Thorsson, V. (2006). The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, vol. 7, no. 5, p. R36.

[7] Boubekeur, S., Camougrand, N., Bunoust, O., Rigoulet, M. and Guerin, B. (2001). Participation of acetaldehyde dehydrogenases in ethanol and pyruvate metabolism of the yeast saccharomyces cerevisiae. *European Journal of Biochemistry*, vol. 268, no. 19, pp. 5057–5065.

[8] Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, vol. 16, no. 9, pp. 575–577.

[9] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. (2009). Blast+: architecture and applications. *BMC bioinformatics*, vol. 10, no. 1, p. 421.

[10] Carter, S.L., Brechbühler, C.M., Griffin, M. and Bond, A.T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, vol. 20, no. 14, pp. 2242–2250.

[11] Cazals, F. and Karande, C. (2008). A note on the problem of reporting maximal cliques. *Theoretical Computer Science*, vol. 407, no. 1, pp. 564–568.

[12] Chen, T., He, H.L. and Church, G.M. (1999). Modeling gene expression with differential equations. In: *Pacific symposium on biocomputing*, vol. 4, p. 4.

[13] De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, vol. 9, no. 1, pp. 67–103.

[14] DâĂŹhaeseleer, P., Liang, S. and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, vol. 16, no. 8, pp. 707–726.

[15] Edgar, R., Domrachev, M. and Lash, A.E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, vol. 30, no. 1, pp. 207–210. Data accessible at NCBI GEO database, accession GEO 21187.

[16] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868.

[17] Fisher, R.A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94.

[18] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620.

[19] Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004). affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, vol. 20, no. 3, pp. 307–315. ISSN 1367-4803.

[20] Goldsmith, Z.G. and Dhanasekaran, N. (2004). The microrevolution: applications and impacts of microarray technology on molecular biology and medicine (review). *International journal of molecular medicine*, vol. 13, no. 4, p. 483.

[21] Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, vol. 10, pp. 1469–1484.

[22] Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic modelsâĂŤa review. *Biosystems*, vol. 96, no. 1, pp. 86–103.

[23] Heinisch, J. (1986). Construction and physiological characterization of mutants disrupted in the phosphofructokinase genes of saccharomyces cerevisiae. *Current genetics*, vol. 11, no. 3, pp. 227–234.

[24] Heller, M.J. (2002). Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, vol. 4, no. 1, pp. 129–153.

[25] Heyer, L.J., Kruglyak, S. and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, vol. 9, no. 11, pp. 1106–1115.

[26] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, vol. 4, no. 2, pp. 249–264.

[27] Laschober, G.T., Ruli, D., Hofer, E., Muck, C., Carmona-Gutierrez, D., Ring, J., Hutter, E., Ruckenstuhl, C., Micutkova, L. and Brunauer, R. (2010). Identification of evolutionarily conserved genetic regulators of cellular aging. *Aging cell*, vol. 9, no. 6, pp. 1084–1097.

[28] Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, L., Chen, L., Chen, T.M., Chin, M.C., Chong, J., Crook, B.E., Czaplinska, A., Varnam, L.R., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T.A., Donelan, M.J., Dong, H.-W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B.A., Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R., Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R.A., Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramee, A.R., Larsen, K.D., Lau, C., Lemon, T.A., Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng, L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S.A., Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N.V., Sivisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K.A., Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpf, K.-R., Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., Wood, M., Yaylaoglu, M.B., Young, R.C., Youngstrom, B.L., Yuan, X.F., Zhang, B., Zwingman, T.A. and Jones, A.R. (2006). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, vol. 445, no. 7124, pp. 168–176.

[29] Li, Z., Li, P., Krishnan, A. and Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, vol. 27, no. 19, pp. 2686–2691.

[30] Lin, S.-J., Kaeberlein, M., Andalis, A.A., Sturtz, L.A., Defossez, P.-A., Culotta, V.C., Fink, G.R. and Guarente, L. (2002). Calorie restriction extends *Saccharomyces cerevisiae* lifespan by increasing respiration. *Nature*, vol. 418, no. 6895, pp. 344–348.

[31] Lin, S.S., Manchester, J.K. and Gordon, J.I. (2001). Enhanced gluconeogenesis and increased energy storage as hallmarks of aging in saccharomyces cerevisiae. *Journal of Biological Chemistry*, vol. 276, no. 38, pp. 36000–36007.

[32] Merry, B. (2002). Molecular mechanisms linking calorie restriction and longevity. *The international journal of biochemistry & cell biology*, vol. 34, no. 11, pp. 1340–1354.

[33] Oates, C., Hill, S. and Mukherjee, S. (2012). On the relationship between odes and dbns. *arXiv preprint arXiv:1201.3380*.

[34] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, vol. 27, no. 1, pp. 29–34.

[35] Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, vol. 18, no. 4, pp. 546–554.

[36] Papoulis, A. and Pillai, S.U. (2002). *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.

[37] Pearson, K. (1896). Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, vol. 60, no. 359-367, pp. 489–498.

[38] Petti, A.A., Crutchfield, C.A., Rabinowitz, J.D. and Botstein, D. (2011). Survival of starving yeast is correlated with oxidative stress response and nonrespiratory mitochondrial function. *Proceedings of the National Academy of Sciences*, vol. 108, no. 45, pp. E1089–E1098.

[39] Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, vol. 2, no. 6, pp. 418–427.

[40] Ristow, M. and Zarse, K. (2010). How increased oxidative stress promotes longevity and metabolic health: The concept of mitochondrial hormesis (mitohormesis). *Experimental gerontology*, vol. 45, no. 6, pp. 410–418.

[41] Sakamoto, E. and Iba, H. (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming. In: *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, pp. 720–726. IEEE.

[42] Scutari, M. (2010). bnlearn: Bayesian network structure learning. *R package*.

[43] Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423,623–656.

[44] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, vol. 13, no. 11, pp. 2498–2504.

[45] Team, B.C. (2002). Bioconductor-open source software for bioinformatics.

[46] Team, R.D.C. (2005). R: A language and environment for statistical computing.

[47] van Dongen, S.M. (2000). Graph clustering by flow simulation.

[48] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Catabolism: Energy release and conservation. In: *Prescott's microbiology*, chap. 10. McGraw-Hill.

[49] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Microbial genomics. In: *Prescott's microbiology*, chap. 16. McGraw-Hill.

[50] Yaramakala, S. (2004). *Fast Markov blanket discovery*. Ph.D. thesis, Iowa State University.

[51] Yeung, K.Y. and Ruzzo, W.L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, vol. 17, no. 9, pp. 763–774.

[52] Yizhak, K., Gabay, O., Cohen, H. and Ruppin, E. (2013). Model-based identification of drug targets that revert disrupted metabolism and its application to ageing. *Nature communications*, vol. 4.

[53] Zou, M. and Conzen, S.D. (2005). A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, vol. 21, no. 1, pp. 71–79.

Chapter 5

Local Non-Parametric Bayesian Clustering Driven Community Detection

5.1 Introduction

Large scale high throughput experiments such as microarrays provide us with information on the gene expression activity that occurs in a cell [55]. Analysing these datasets allows us to potentially gain a better understanding of the fundamental processes that occurs within a cell. Patterns uncovered from the data can be used to generate hypotheses regarding the combinatorial interactions that occur between and amongst genes and gene products.

These microarray experiments involve a large number of variables in the form of probe sequences, with the expression of these sequences studied under various biological conditions [31]. Sequences are grouped together in sets, with each of these probesets associated with a set of genes. Multiple of these probesets can be associated with a single gene and multiple genes can be associated with a single probeset. Where expression generally refers to the hybridization of extracted and processed mRNA.

Each of the biological conditions is represented by a microarray chip, with the extracted and processed fluorescent mRNA hybridized to the chip. The observed intensity values are then analysed to produce a vector of values. The intensity values serve as a proxy for expression of a probeset, and therefore genes, under the particular biological conditions. Several factors contribute to the fact that not many biological conditions are studied during these experiments, thus leading to a large number of variables observed under a small number of conditions [38, 31, 55].

Classical statistical methods, such as the t-test often used to determine differential expression, frequently lack sufficient statistical power to draw meaningful conclusions from these experiments [5]. A general approach to study

these datasets involves some form of dimensional reduction or clustering [27, 50]. This approach allows the study of sets of genes, which may capture information that may be missed when studying genes in isolation [49]. From these clusters or sets, putatively biologically meaningful subsets can be produced and studied. Studying sets of genes may improve our understanding of the fundamental processes that govern cell function and development. From a computational point of view these sets may also have dimensions more favourable for various statistical techniques.

An alternative approach to improve the descriptive capacity of various mathematical and statistical models that are applied to these experiments, is to include additional information in the model [47]. In well studied organisms additional information is much easier to obtain, however, in other organisms and often in the context of exploratory analysis and hypothesis generation, additional information is not readily available [56].

A modelling framework that naturally includes additional information is that of Bayesian models. In these models a full probability model is specified where variables are considered to be observations from a probability distribution. These models use conditional probabilities to include prior information on the variables and then model how these prior beliefs are altered when data is observed [18]. Bayesian models have been used extensively in the modelling of gene expression and microarray data. With a well chosen prior, these models have shown good results in the context of limited data [48, 42, 32].

Here we applied a non-parametric Bayesian variable selection model, in the form of an Indian Buffet Process, which produced subsets of variables [21]. The probabilistic association between these variables was then assessed using a non-parametric Bayesian mixed modelling approach, in the form of a Dirichlet Process Mixture model [28]. The resultant probabilities were viewed in the context of a graph, whereby we determined the path that has the highest probability connecting all nodes, the maximum spanning tree [29]. To produce putatively biologically meaningful subsets from this tree we investigated communities estimated from the tree structure. These communities served as putatively interacting genes generated from a probabilistic framework given a local background. This exploratory approach was applied to time-series gene expression data for *Saccharomyces cerevisiae*, with the objective of proposing sets from a probabilistic context for the purposes of generating putatively biologically relevant hypotheses. The resulting network was visualized using Cytoscape [45].

5.2 Results and Discussion

We applied our exploratory analysis approach to a time series microarray dataset obtained from NCBI [13].The experiment involved the study of non-caloric restricted *Saccharomyces cerevisiae* cells that had been immobilized

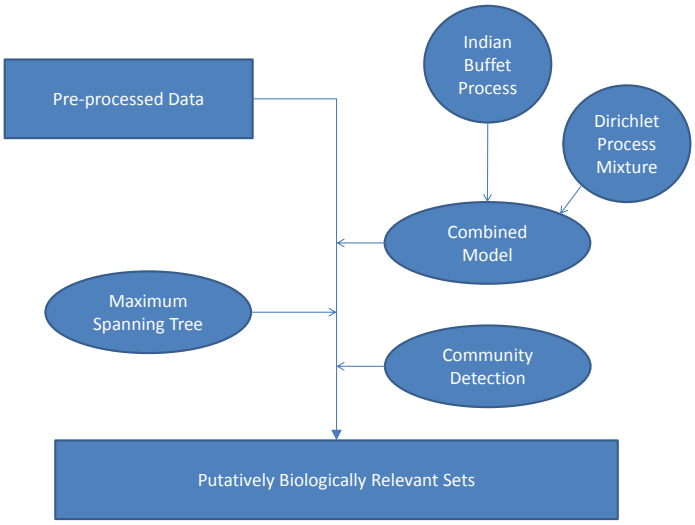


Figure 5.1: Methodology

over a period of 17 days. Samples of mRNA were taken at day 1,3,5,10 and 17 respectively, with sampling done in triplicate. The raw expression data was obtained and normalized, resulting in an expression matrix of $\log_2$ intensity values.

To generate the putative sets, we processed our data by normalizing the rows of the expression matrix, obtained from a microarray experiment. This was done by dividing the respective rows by the row sum of the intensity values. Therefore, we effectively modelled relative intensity values.

This method was then applied to determine the probability of probeset interactions, by the proportion of probeset co-occurrence in a locally generated cluster. The local background for these clusters was obtained from an Indian Buffet Process, with the clustering performed using a Dirichlet Process Mixture Model. The number of times probesets co-occurred in the same cluster based on samples from a truncated Markov Chain Monte Carlo sampling scheme was determined, which was then normalized by the number of samples resulting in a proportion of co-occurrence.

The predominant trend throughout this probabilistic association was obtained using a maximum spanning tree. From this tree we generated sets by determining the community structure present in the topology of the tree.

These communities estimated from the tree structure were therefore our hypothesized biologically relevant sets estimated from a probabilistic context. The methodology is outlined in Figure 5.1.

.

The approach highlighted above was applied to a subset of the microarray dataset, where we only model probesets that match genes that occur in the metabolic pathway of *Saccharomyces cerevisiae* as determined by the Kyoto

Encyclopedia of Genes and Genomes (KEGG) [35]. The probeset to gene matching was determined using a BLAST comparison [1], for which we used the offline algorithm provided by NCBI [9]. We performed the majority of our analysis on probeset and only translated these probesets to genes when performing biological interpretation.

Our reduced expression matrix consisted of 743 probesets and 5 time points, with 3 repeats per time point. We analysed the expression matrix after taking the average over the repeats. As discussed in the Materials and Method section, a truncated approximation approach was taken with the Markov Chain Monte Carlo (MCMC) sampling, for this a $T = 20$ was deemed adequate. Convergence of the MCMC sampling chain was assessed using trace plots, autocorrelation plots and the Geweke statistic [19]. The algorithm and the convergence diagnostics were all implemented in Python with the use of the PyMC package [39, 43]. Convergence was assumed to occur based on the combination of the trace plots, autocorrelation plots and the plots from the Geweke analysis. For the parameters investigated the autocorrelation plots appeared to decay exponentially for most parameters, while the trace plots remained relatively stable. The parameters produced a majority of z-score values within two standard deviations from the mean, these parameters are thus within an approximated 95 percent confidence interval around the parameter's sample mean. A few trace, Geweke plots and autocorrelation are presented in Figures 5.2 and 5.3 as examples. The MCMC sampling scheme consisted of 97000 iterations with a burn in phase of 87000, thus resulting in a total of 10000 iterations sampled. It was run with different starting values for the parameters generating 2 separate chains, each suggesting convergence.

The run time of our method for increasing numbers of iterations was also assessed and the results are presented in Figure 5.4. From this it appears that the run time of the method is linear in the number of iterations for longer iterations. The proportionally slower run time for lower iterations may be the result of a lack of convergence. The method is currently implemented in series, but is quite well suited to implementation in parallel and could theoretically be implemented on multiple machines without the need for communication between these machines, thus potentially leading to a more efficient memory and time footprint [34]. There are also alternative MCMC sampling schemes that could be used to improve the ability of the method to explore the posterior landscape, approaches such as slice sampling [52].

The resultant maximum spanning tree initially consisted of probesets (see Figure 5.5). Probeset to gene ambiguity was subsequently taken into account in order to generate a gene-based representation (see Figure 5.6). There was a total of 73 communities estimated from the probeset representation of the weighted tree structure using the Python implementation of the Louvain Method [7]. The edge weights were the respective probabilities of the probesets co-clustering. There communities were then pruned, only retaining communities that consisted of more than 20 elements. This produced a total of 8

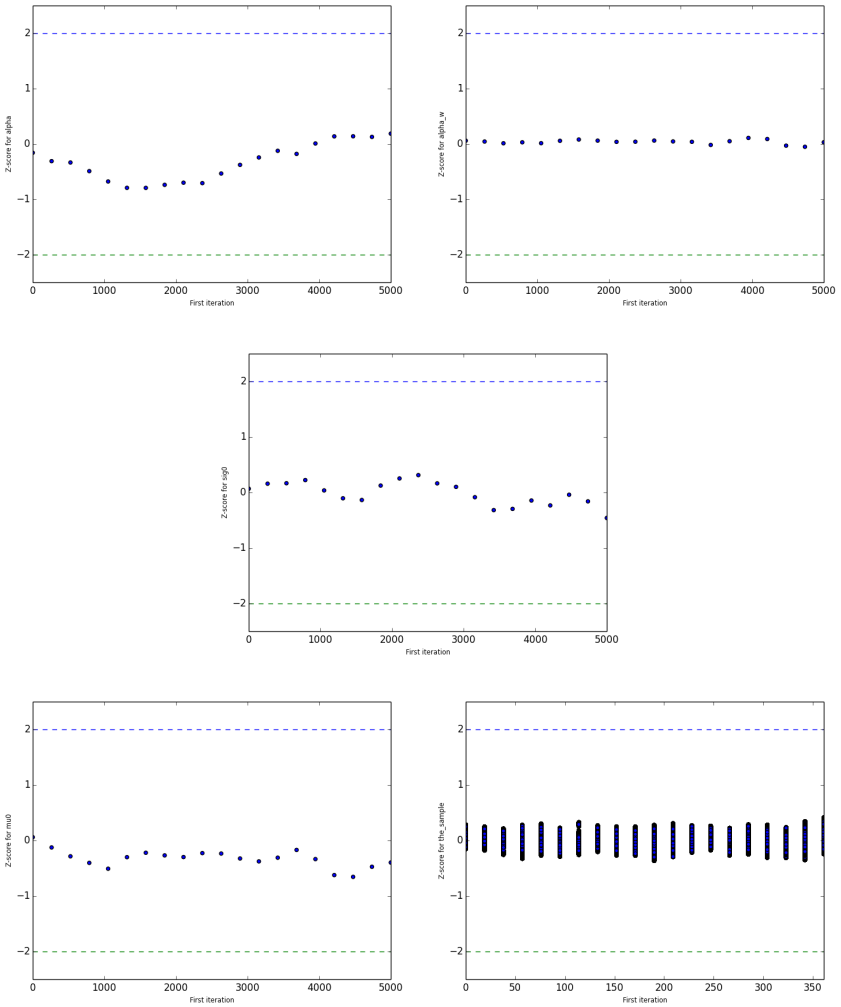


Figure 5.2: Geweke Plots of Z-scores

The above are Geweke plots of the z-scores resulting from comparing the means and variances of the first 10 percent of a chain with the means and variances of the last 50 percent of a chain. Chains represent intervals from the MCMC run. If the chain has converged then the majority of z-score values should fall between -2 and 2. The first plot is for α, which is the hyperparameter of the Dirichlet Process Mixture. The second plot is for α_w, which is the hyperparameter of the Indian Buffet Process. The third plot is for μ_0 and the forth is for σ_0, these are the hyperparameters of the Log-Normal Base Distribution and the final plot is for the matrix of random variables indicating the Hadamard product of C and Z.

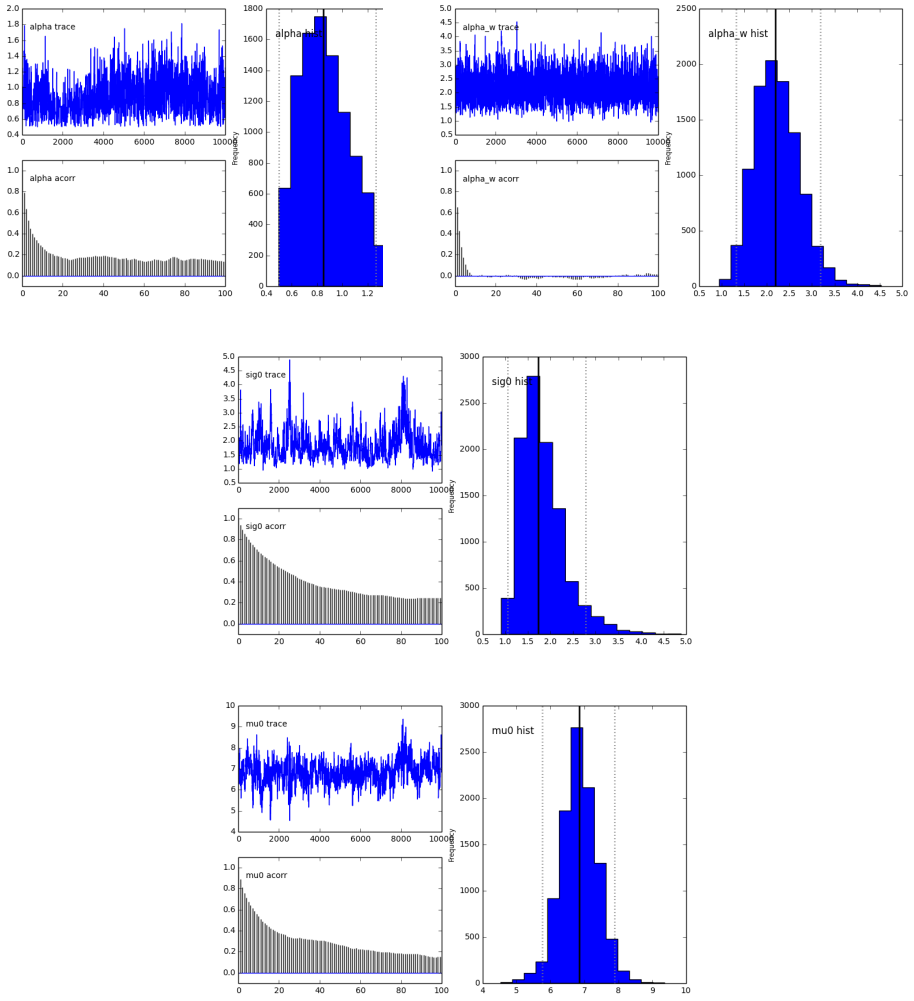


Figure 5.3: Trace Plot and Autocorrelation of MCMC: 10000 samples

Trace and autocorrelation plots of a few parameters sampled from the posterior using MCMC. The plots are given in groups, with the trace plot in the upper left, the autocorrelation plot on the lower left and a histogram of parameter values on the right. These group plots represent the following parameters in order: first α the hyperparameter of the Dirichlet Process Mixture, second α_w the hyperparameter of the Indian Buffet process, thirdly μ_0 and finally σ_0 which are both the hyperparameters of the Log-Normal Base Distribution. The trace plots for μ_0, σ_0, α and α_w are relative stable. The autocorrelation for μ_0, σ_0 and α_w decay to zero for larger lag values. It is generally preferred to have the trace plots converge, oscillate around a particular value without jumping, and the autocorrelation to exponentially decay quickly to zero.

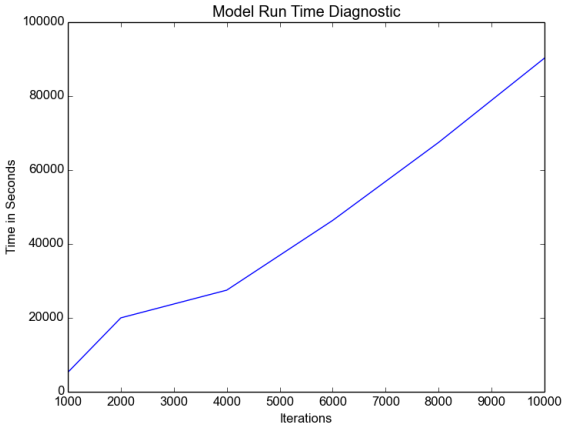


Figure 5.4: Run time of Algorithm

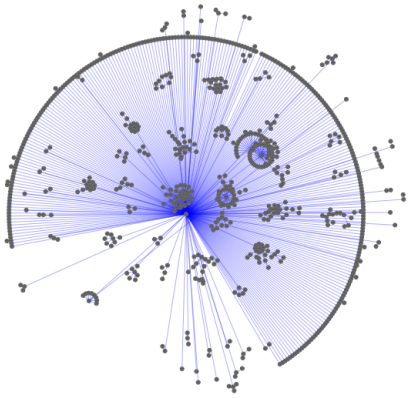


Figure 5.5: Inferred Tree Structure: Probesets

communities. The biological relevance of these communities were determined using Gene Ontology (GO) enrichment and also by investigating the most connected node in the community. The communities were also translated into a gene representation and projected onto the corresponding tree structure.

5.2.1 Gene Product Characterization and Annotation

In order to functionally characterize and annotate the set of genes uncovered from the topological community structure resulting from our method, we performed GO-enrichment on each of these sets, utilizing GOEAST [57]. The approach used by this application to determine the significant association of terms to the genes is discussed in the Materials and Methods section.

From the network, the largest community detected consisted of 243 probesets. The most connected node in that community corresponded to the center node in Figure 5.6. Evidence suggesting this node's ($YKL085W$) involve-

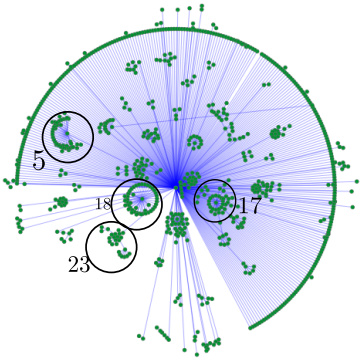


Figure 5.6: Inferred Tree Structure:Genes

The resultant maximum spanning tree translated from probesets to genes. The encircled areas indicate the topological location of the respective communities inferred from the weighted tree structure, the edge weights are the probabilities of co-occurrence of probesets in the same cluster.

ment in yeast cell ageing was presented by [12]. Its immediate neighbours included $YDL140C$, $YDR404C$ and $YNL229C$ which are all known regulators of $YKL085W$ [10].

The majority of significant communities are immediate neighbours of $YKL085W$, where we referred to the node connecting the respective community structure to $YKL085W$ as the *center node*. These *center nodes* also had the highest degree in their respective communities. A few of the results from the encircled communities are discussed below, with a list of the genes in the respective communities discussed given in Table 5.1.

Community 18

This community was enriched for *amino acid metabolic processes*. This term encompasses chemical reactions in a cell involving amino acids or carboxylic acids. The putative role that amino acids play in chronological ageing of yeast cells was discussed in [3]. Amino acids can perform the role of aerobic carbon source, where it was proposed in [8] that areobic growth before the non-dividing stage may promote longevity. The recycling of amino acids may also play a role together with mitochondrial damage and autophagy to effect the chronological age of non-dividing (stationary-phase) yeast cells [3] where autophagy refers to the process of self-catabolism.

Alternative theories of chronological ageing, such as caloric-restriction-based theories, have also suggested that amino acids play an important role in ageing [40]. There is also evidence to suggest that the restrictive intake of amino acids produces conditions similar to those produced by applying caloric restrictions [37].

Table 5.1: Genes found in Communities

Community Name	Gene List
Community 5	YBR026C YBR277C YBR278W YDL045C YDL093W YDR037W YDR483W YER003C YER183C YFR019W YFR052C-A YFR053C YGL125W YGL256W YGR088W YGR199W YIL014W YIL139C YIL160C YIR019C YIR032C YJL140W YJL167W YJR103W YLR180W YLR382C YMR099C YMR246W YNL048W YNL331C YNR012W YOR109W YPL028W YPL268W YPR187W
Community 17	YPL167C YLR450W YLR432W YNL220W YGR043C YHL032C YGR192C YGL055W YHR216W YDR242W YNL151C YJL216C YER005W YNR008W YER087W YKL150W YDR441C YPR145W YCR005C YCR107W YNL267W YAR075W YCL052C YOL097C YIR037W YJR009C YGR110W YIL094C YGR012W YER087C-A YHR137W YOR360C YOR202W YIL125W YHL011C YJL052W
Community 18	YAL037C-B YAL038W YCL004W YDL021W YDR234W YDR294C YER043C YFL022C YFR025C YGL001C YGL202W YGL257C YIL162W YIL172C YIR034C YJL130C YJL221C YJR006W YJR139C YJR153W YKL001C YKL045W YKL141W YKL184W YKR089C YLL041C YLR028C YLR100W YLR101C YML022W YML106W YMR113W YMR250W YMR300C YNR043W YOL059W YOL140W YOL157C YOR143C YPL097W YPL160W YPR006C YPR060C YPR113W YPR127W YPR167C
Community 23	YPR140W YPR062W YOR176W YOR121C YOR120W YNR016C YNL037C YNL009W YML056C YLR354C YJR109C YJR024C YGR292W YFR047C YFL053W YDR437W YDR148C YDR001C YCL050C YBR299W YBR084W YBR006W YBL068W YBL015W

Community 23

Similarly to community 18, this community was enriched in the *carboxylic acid metabolic process*, which is directly connected to *amino acid metabolic processes*. More specifically, community 23 was enriched for *dicarboxylic acid metabolic processes*. It was proposed in [54] that expression of *Indy*, a gene closely related to dicarboxylate cotransporter, has an important effect on longevity in *Drosophila*.

This community was also enriched for the *tricarboxylic acid cycle* (TCA). It was highlighted in [3], that retrograde signalling, the signalling that occurs between mictochondrion and nucleon, may potentially influence cellular longevity. Retrograde signalling is primarily responsible for carbon flow regulation using the TCA cycle.

The center node for this community is $YNR016C$, also known as *Acetyl-CoA carboxylase* (ACC1). A biological function attributed to ACC1 is histone acetylation and it has been established that SNF1 has a potential inhibitory effect on ACC1 [15, 46]. The activity of SNF1 in aged cells in the presence of abundant nutrients, namely glucose, is discussed in [24]. Furthermore a putative link between histone acetylation, mediated by SNF1 inhibition, on longevity is discussed in [30]. This was not specifically linked to ACC1, though this may indicate a valid avenue for further investigation.

Community 5

Probesets in this community were enriched for *protein o-linked glycosylation*. In [6] it was postulated that *protein o-linked glycosylation* may be involved in a nutrient signalling capacity. This would putatively agree with the experimental conditions underlying our dataset, namely that of excess nutrient environment.

Community 17

The *center node* of this community is $YNL267W$ also known as *Phosphatidylinositol 4-Kinases* (PIK1) [10]. This gene plays an essential role in the nucleus of yeast [16]. Evidence presented by [53] suggests that $PIK1$ is involved with autophagy and arguments proposed by [2] highlight the putative role that autophagy plays with regards to cellular longevity during the stationary-phase.

5.3 Materials and Methods

Dirichlet Process Mixture Model

A mixture model is a model that is built on the underlying assumption that the data is generated from a mixture of functions, such as probability distributions. Each variable is a realization from some probability distribution.

Clusters can be obtained from this. The variables that have the same probability distribution are assumed to belong to the same cluster.

A Dirichlet Process Mixture model is a mixture model where the underlying generative distributions have their parameters drawn from the samples of a Dirichlet Process [14]. A Dirichlet Process is a stochastic process that forms part of the class of non-parametric Bayesian models. The concept of non-parametric in terms of a Bayesian model refers to a model that is 'infinitely' parametric, in other words an infinite number of parameters for the particular distribution.

In particular, the Dirichlet Process is a distribution over probability distributions and samples from the process are discrete probability distributions with the same support as a given base distribution, denote the base distribution by H, which is one of two parameters that define the process. We denote a sample from the Dirichlet Process, thus a discrete distribution, by G. The other parameter of the Dirichlet Process is known as a concentration parameter, denoted α. It controls the concentration of the mass of the sampled distribution such that as $\alpha \to \infty$, $G \to H$ pointwise.

The formal definition of a Dirichlet Process is given as:

Definition 5.1. Dirichlet Process Let $\alpha \in R^+/0$ and let H be some probability distribution (or density) with support A. Then $G \sim DP(H, \alpha)$ if for any finite set of partitions of A, say $A_1 \cup A_2 \ldots \cup A_k = A$, we have that

$$(G(A_1), G(A_2), \ldots, G(A_k)) \sim Dirichlet(\alpha H(A_1), \alpha H(A_2), \ldots, \alpha H(A_k)) \quad (5.3.1)$$

This definition is not constructive, since it gives no indication of how to construct a Dirichlet Process to obtain G. There are several alternative representations of this process, in particular we utilized the stick-breaking representation proposed by [44] to construct the Dirichlet Process used in our method. This representation is given mathematically by the following set of equations:

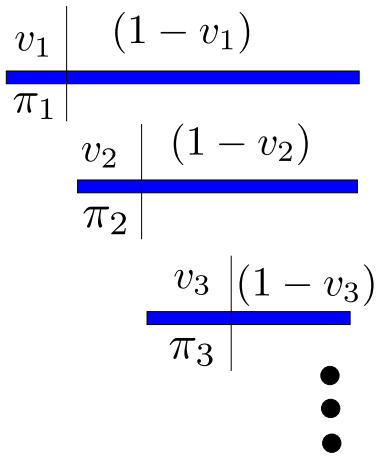
$$\begin{aligned} k &= 1, 2 \ldots, \infty \\ \theta_k &\sim H \\ v_k &\sim Beta(1, \alpha) \\ \pi_k &= v_k \Pi_{l=1}^{k-1}(1 - v_l) \\ G(A) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(A) \end{aligned} \quad (5.3.2)$$

Here $\theta_k \in A$ is the k^{th} sample from the base distribution H with support A, with the support of a probability distribution is the set of values for which it is defined. Further, v_k is the k^{th} draw from a Beta Distribution and π_k is a probabilistic weight assigned to the support value θ_k, such that $\sum_{k=1}^{\infty} \pi_k = 1$. We can see from the equation of G, that G is a countably infinite atomic

distribution with atoms given by $\delta_{\theta_k}()$, where $\delta_{\theta_k}()$ is the dirac delta measure, such that $\delta_{\theta_k}(\theta_k) = 1$ and $\delta_{\theta_k}(\cdot) = 0$, otherwise. We note from the above that $\sum_{k=1}^{\infty} \pi_k = 1$, therefore we can say the π_k represents that weight, or probability mass, associated with the atom $\delta_{\theta_k}()$.

Intuitively we can view the above process as the breaking of and discarding pieces of a unit stick. First a support value is sampled from the base distribution, then a location to break the unit stick is sampled from a Beta Distribution. Then the part we broke off is taken and its length is assigned as a probability associated with the support value. The process is repeated, sampling another support value and then breaking parts off from what remains of the unit stick. This process is continued for an infinite number of iterations. By continually breaking from the remains of the unit stick we are assured that the probability mass assigned to the support values will asymptotically sum to 1. This process is illustrated in Figure 5.7. It was shown by [44] that G generated from the above algorithm follows a Dirichlet Process.

Figure 5.7: Stick Breaking Representation: Dirichlet Process



In terms of the objective of clustering, we interpreted π_k as the probability of assigning a variable, or object of interest, to a cluster k. It is worth noting that the support value θ_k is not restricted to a scalar, therefore θ_k can indicate a set of parameters associated with the cluster k. These parameters can then in turn indicate a distribution for a particular cluster k. The stick breaking process above therefore provided a model to probabilistically pick an infinite set of parameters for some distribution. Given the infinite set of assignment

probabilities, cluster assignment is independent of each other. This implies a conditional independence given by:

$$P(c_i, c_j|\pi_1, \pi_2, \dots) = P(c_i|\pi_1, \pi_2, \dots)P(c_j|\pi_1, \pi_2, \dots),$$

which applies to all clusters, where a cluster is indicated by c with a subscript. This therefore implies that the cluster assignment for, say N, variables or objects follows a Multinomial Distribution:

$$(c_1, c_2, c_3, \dots, c_N) \sim Multinomial(\pi_1, \pi_2, \dots, \pi_N) \tag{5.3.3}$$

Together the set of equations 5.3.2 and the equation in 5.3.3, provided us with a non-parametric prior over cluster assignments and parameter values. The observations of an expression matrix are continuous values, while the realization of the distributions sampled from the Dirichlet Process are discrete. Therefore we assumed that the realization of these sampled distributions are actually parameters of a Beta Distribution. Thus our expression matrix, once normalized, was assumed to be generated from a mixture of Beta Distributions, where each variable is a set of realization from one of the infinite number of Beta Distributions:

$$y_i|\theta, c_i \sim Beta(\theta_{c_i}) \tag{5.3.4}$$

It is important to note that we have continuously referred to θ as a support value, which in our case referred to the parameters of a Beta Distribution. This support value therefore has domain $R^+ \times R^+$ and is thus a two element vector with each element having support over the positive real numbers. The support of the base distribution therefore has to be the same. We took an independent bivariate Log-normal Distribution as our base distribution. The probability density of a Log-normal Distribution is given in 5.3.5,

$$f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\{-\frac{(lnx - \mu)^2}{2\sigma^2}\}; x > 0, \tag{5.3.5}$$

Finally we defined hyperpriors for the prior parameters α, μ and σ. The hyperpriors where chosen to be uninformative. We took α to be a realization from a Uniform Distribution, μ_j a realization from a Normal Distribution and σ_j to be a realization from a Uniform Distribution, for $j = 1, 2$ which denoted the two respective parameter sets for the bivariate Log-normal Distribution. The hyperpriors for the base distribution was chosen to represent a wide range values to capture the notion uninformative belief surrounding the parameter values of the respective Beta Distributions.

$$\alpha \sim Uniform(0.5, 10)$$
$$\mu_j \sim N(0, 100)$$
$$\sigma_j \sim Uniform(0, 100)$$

We took the hyperprior for σ_j to be uniform rather than the classical Gamma Distribution generally used, it has been argued in [17] that uniform prior is preferred when working with hierarchical models. Further we assumed equal means and variances for the independent bivariate Log-normal Distributions, this implied

$$\mu_1 = \mu_2$$
$$\sigma_1 = \sigma_2,$$

with the corresponding co-variance elements equal to zero.

Our mixture model is therefore fully specified. The graphical depiction of our model is given in Figure 5.8. This uses the plate representation where all variables in the same plate are indexed over the same indices.

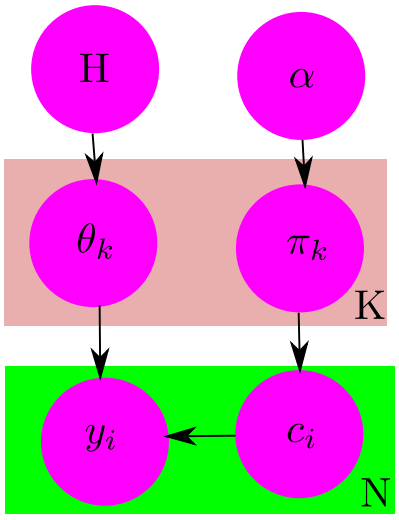


Figure 5.8: Graphical Plate Model: Dirichlet Process Mixture Model

5.3.1 Indian Buffet Process

Genes interact together in modules and sets, often in a combinatorial manner. There is no guarantee that genes that do interact together have the same expression pattern. The patterns between these genes may be highly correlated, inversely correlated or may show signs of little to no correlation. We modelled this concept from an exploratory analysis approach. We therefore proposed sets of genes that are then further analysed in terms of their expression patterns, thus capturing these patterns given a local background.

To initially generate a subset of genes, we viewed the problem in the context of binary classification of variables given a set of features. A feature describes an attribute of a variable. In our case variables are probesets and features represent some arbitrary attribute, left unspecified. For a particular feature, the genes that were associated with this feature were viewed as a putative gene set. The concern was then defining features to pick, and how many of these features we should use. We used an Indian Buffet Process to solve this, which uses an infinite number of independent features [21].

A Indian Buffet Process defines a prior over equivalent classes of binary matrices. The original representation of this process defined it as the limit of a prior over equivalent classes of finite binary matrices, where the limit is taken over the number of columns [21]. A binary matrix is an $N \times K$ matrix, say Z, whose elements are either 0 or 1. The N rows indicate the variables that we are interested in, the K columns are the independent features with $K \to \infty$ and the elements of the matrix indicate if there is an association between the variable and the feature, in the case of such an association the corresponding element will be 1.

We utilized the stick-breaking representation proposed by [51], instead of the originally proposed representation. The stick-breaking process was shown to be equivalent to the process defined above. This stick-breaking process is given by the following set of equations, where $i = 1, 2, \ldots, N$ and $k = 1, 2, 3, \ldots, \infty$:

$$
\begin{aligned}
w_k &\sim Beta(\alpha_w, 1) \\
b_k &= \prod_{l=1}^{k} w_l \\
z_{ik}|b_k &\sim Bernoulli(b_k)
\end{aligned}
$$

The above can be intuitively understood as the iterative breaking of pieces from a unit stick. For the first iteration we sample, from a Beta Distribution, a proportion w_1 of the unit stick we will break off and record the length of the part we broke off, b_1, discarding the rest of the stick. Then we use the calculated length, as a probability, to generate a vector of N binary values. The second iteration then involves sampling another proportion, w_2 from a

beta distribution, but this time the proportion is in terms of the piece we previously broke off, recording its length, $b_2 = w_2 * b_1$, and then sampling another binary vector. This process continues infinitely, each time breaking off a proportion of the stick that we have broken off previously. This is illustrated in Figure 5.9. A plate model graphical representation of the above process is shown in Figure 5.10.

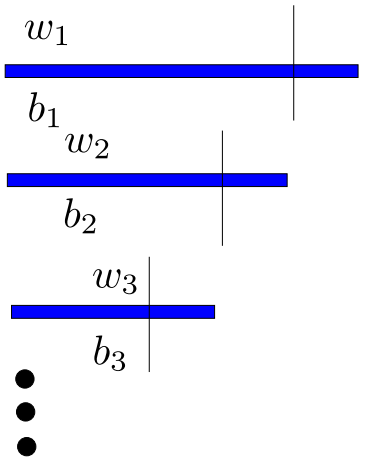


Figure 5.9: Stick Breaking Representation: Indian Buffet Process

Asymptotically, this process will generate a binary matrix with infinite columns where the probabilities associated to the Bernoulli Distribution exponentially decay as more features are sampled. This leads to the number of columns that have no non-zero element being finite. The distribution of the number of columns with at least one non-zero element has been shown to follow a $Poisson(\alpha \sum_{i=1}^{N} \frac{1}{i})$ distribution [51, 21].

5.3.2 Combined Model and Truncated MCMC Sampling

To generate our putatively biologically relevant sets, we determined the probabilistic association between genes given a local background. This was done by clustering a subset of genes. The subsets were generated from each of the columns of the binary matrix produced by the Indian Buffet Process. The clustering was performed on this subset, by a Dirichlet Process Mixture Model. The probabilistic association was then finally determined by the normalized

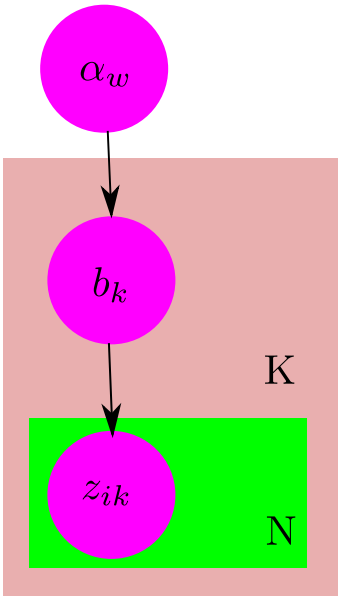


Figure 5.10: Graphical Plate Model: Indian Buffet Process

frequency of genes clustered together. This required that we sample clusters from the posterior distribution of our model. Therefore, Markov Chain Monte Carlo (MCMC) simulations were used to sample these cluster probabilities and the respective mixture parameters associated with the clusters [23, 20, 11].

The combined model of the two processes is graphically presented in Figure 5.11. From the Indian Buffet Process we sampled an NxK binary matrix, Z, and from the Dirichlet Process we sampled a K x 2 matrix, θ, which contains parameter values for a Beta Distribution and a N x K matrix, C, of cluster identities. We then took the Hadamard product of C and Z to determine the variables to be analysed for each k, say N_k^*, and their respective cluster identities for each of the individual features, obtained from a Multinomial Distribution. For each of these features we applied the likelihood, the Beta Distribution with the parameters given by the cluster identity's corresponding row in θ. With the likelihood it was assumed that the variables, which are vectors of normalized expression values, are independently distributed given the cluster assigned parameter values and features.

In order to generate samples of cluster assignments from the posterior of our model we truncated the column space of the binary matrices and the dimension of the distribution sampled from the Dirichlet Process. A finite dimensional approximation was therefore applied to both the Indian Buffet Process and the Dirichlet Process. The concern therefore became choosing an adequate truncation value, T, such the model is computationally feasible, but still expressive enough. In the case of the Dirichlet Process, [25] initially

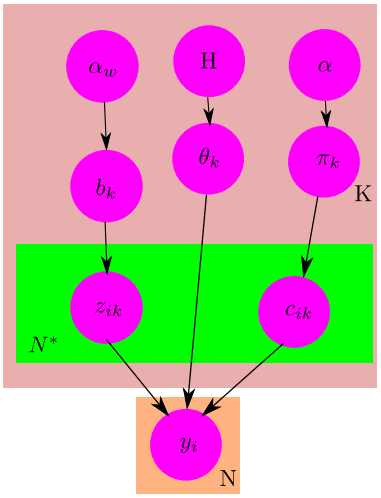


Figure 5.11: Graphical Plate Model: Combined Model

proposed this approach and [36] provided a way to obtain T such that the expected final probability mass assigned to π_T is arbitrarily small, say ϵ. This will ensure that we have modelled a large portion of the probability mass:

$$T \approx 1 - \alpha log(\epsilon) \tag{5.3.6}$$

For the Indian Buffet Process, by observing the stick-breaking representation, there are notable similarities with the stick-breaking representation of the Dirichlet Process, with the Dirichlet Process we continually broke pieces off from what remained of the unit stick, while with the Indian Buffet Process we continually broke pieces from the piece that was broken off. Given this similarity it was argued in [51] that the truncation approximation can also be applied to the Indian Buffet Process. To determine the truncation value T, for the Indian Buffet Process we tried to ensure that the expected probability assigned to b_T is again arbitrarily small, say equal to ϵ. The expectation of b_1 is given by $\frac{\alpha_w}{1+\alpha_w}$, which follows from the beta distribution. Each of the b_k random variables, for $k = 1, 2, 3, \ldots$, are independent, thus the expectation of b_T is given by $E(b_T) = (\frac{\alpha_w}{1+\alpha_w})^T$. We therefore have that T is approximated as given by the set of equations 5.3.7,

$$\begin{aligned} \epsilon &= (\frac{\alpha_w}{1+\alpha_w})^T \\ \log(\epsilon) &= T\log(\frac{\alpha_w}{1+\alpha_w}) \\ T &= \log(\epsilon) \div \log(\frac{\alpha_w}{1+\alpha_w}) \\ T &\approx -\alpha_w \log(\epsilon) \end{aligned} \tag{5.3.7}$$

.

In the set of equations above we used the approximation of $\log(\frac{\alpha_w}{1+\alpha_w}) \approx \frac{-1}{\alpha_w}$. Therefore the choice of truncation value with regards to both the Dirichlet process and the Indian Buffet Process is primarily a function of their respective hyperparameters.

To navigate the sample space during the MCMC simulation the Metropolis sampler was applied, where a centered Gaussian Distribution was used for the proposal distribution in the variable-based updates. No block updates were therefore performed, as each random variable was updated individually. This particular choice of a proposal distribution effectively explored the search space with a random walk. This introduced an additional parameter that needed to be estimated, namely the variance of the Gaussian Distribution. If the variance is too large then the space might be sparsely explored, while if the variance is too low then the chain might be slow to converge. The process of choosing an appropriate variance is called tuning. Generally the variance is adjusted such that the acceptance ratio is 0.5 [41]. The acceptance ratio is the ratio of the number of accepted proposed values compared to the number of rejected proposed values for a parameter. We adjusted the standard deviation of the proposal distribution during the burn-in phase of the sampling procedure.

5.3.3 Minimum Spanning Tree

The minimum spanning tree problem is defined as determining a spanning path that has minimum weight from a given graph G, where the graph $G = (V, E)$ is defined by a set of nodes V, and a set of edges $E = \{v_i, v_j\}, v_i, v_j \in V$. A spanning path $p \in P$, where P indicates all possible spanning paths, is a path that contains all nodes in V, such that it consists of an alternating sequence of edges and nodes where no node or edge is repeated. The weight is a positive value that is assigned to a given edge $e \in E$ and a path is said to be of minimum weight if the sum of the weights of all the edges of that path is less than or equal to the sum of the weights of any other path.

The algorithm to determine the minimum spanning tree used here is the Kruskal Algorithm proposed in [29]. The algorithm proceeds in the following steps:

1. Create a set of trees, T, such that each tree is a node from the graph

2. Create a sorted set, say U, with all edges from the graph sorted by weight

3. While the T is not spanning and U is not empty do

   a) Remove the shortest weighted edge from T, mark it as current

   b) If the current edge has nodes in two different trees, merge them. Else discard edge

The result from the above algorithm is a minimum spanning forest, if the result is a set of disconnected trees, otherwise it will be a spanning tree. The complexity of the above algorithm is determined to be $O(|E| \log |E|)$.

If the weights of a graph are inverted in some way, then the resulting tree (or forest) will be a maximum spanning tree. These trees allow a de-convoluted representation of a potentially complex graph structure. When the weights between nodes are probabilities the spanning trees provide a most probable path between the nodes. If nodes are genes (or probesets) this provides an image of the most probable interaction between these genes; given the context of the dataset. The NetworkX package in Python was used to determine the maximum spanning tree [22].

5.3.4 Community Structure Detection

Communities are defined to be densely connected subgraphs in a graph. These topological structures naturally capture the idea of modules of gene interaction, whereby genes interact in a combinatorial set-based manner. Thus we assume that these subgraphs, when applied to a probabilistic association graph derived from gene expression data, captures this dynamic.

Exact solutions to community detection is generally computationally intractable. Therefore most algorithms provide some heuristic approximation, with the objective to generate sets of nodes that are densely connected to each other and sparsely connected to all other nodes. This is measured by the modularity of the network, which provides a value between -1 and 1, where higher values of modularity indicate a greater disparity between the edge density within a community compared to the edge density between communities.

The algorithm used to detect high modularity communities for our given graph structure was the Louvain algorithm proposed in [7]. This approach uses modularity as an objective function in a two phased approach that is repeated iteratively. The first phase is initialized by assigning each node to a separate community. The nodes and their respective neighbours are then investigated and a node is reassigned to the community belonging to its neighbour only if

	Valid ID's in Gene Set	**Valid ID's in Gene Set Complement**	
Genes have GO term	b	d - b	d
Genes don't have GO term	a -b	n+b-a-d	n-d
	a	n-a	n

the change in modularity is a positive maximum. When no switching of communities occurs, the algorithm passes to a second phase. This phase involves constructing a new graph from the previous graph. In this new graph nodes are the communities produced by phase one, with edge weights corresponding to the sum of the weights connecting nodes from the two corresponding communities. Thus there exists an edge between two new nodes in the new graph if there exists at least one edge between a pair of old nodes of the previous graph where each of these old nodes are elements of two communities. The construction of self-loops is also possible during the second phase. When phase two is completed, phase one is applied to this new graph. This process continues for a certain number of passes or until there is no change in the community structure. The resulting nodes after the final second phase then indicate the communities. Different levels of the generated hierarchy can be analysed, to identify potentially meaningful structure related to sub-communities.

5.3.5 Gene Ontology Enrichment

The Gene Ontology (GO) is a project that involves collecting a standardized set of terms that describe gene products and gene annotations. There are several tools available to determine overrepresented GO terms for a given set of genes or probesets. This process is known as GO enrichment. The structure of these terms is hierarchical where the top level of the hierarchy describes more general terms and the terms become more specific and specialized as one descends the hierarchy. This structure reveals the connected nature of different overrepresented terms and how they related to one another.

Various statistical techniques have been proposed to determine if a set or subset of GO terms are enriched for a particular set of genes. The standard test utilized to determine enrichment or significant association is the hypergeometric test. This involves the construction of a contingency table for each GO term.

In the table a is the number of valid id's, or GO-terms, for genes in the set we are investigating and the microarray chip consists of a total of n genes. For a particular GO term there are b number of genes in a associated with it and there are d number of genes in n associated with it. Then for a particular GO term a p-value for the hypergeometric test can be calculated by

$$p = \sum_{x=b}^{d} \frac{\binom{d}{x}\binom{n-d}{a-x}}{\binom{n}{a}} \tag{5.3.8}$$

These p-values are then adjusted using the Benjamini-Yekutieli adjustment, to produce a false discovery rate. This is the approach implemented by GOEAST [57], a web-based tool for GO-Enrichment. The output produced by the method is a tree hierarchy in which significantly enriched GO-terms are highlighted. The tree indicates more general GO-terms at the root and as the tree descends we have less general GO-terms.

5.4 Conclusion and Future Work

Our method involved the application of primarily two non-parametric Bayesian models. The first, an Indian Buffet Process, is seen as a prior on binary matrices. This process was used to select subsets from our data. These subsets were then clustered using a Dirichlet Process Mixture model. By using a sampling scheme, we could then estimate the probability of co-occurrence of our variables. This estimation resulted in a complete graph, that was then reduced using a maximum spanning tree. Community structures present in this weighted tree then allowed us to generate subsets, that may provide us with putative contextual relevance.

In applying our methodology to a subset of a *Saccharomyces cerevisiae* dataset, we uncovered a rich structure. This structure putatively appears to capture the underlying experimental dynamics and postulated communities that hypothesize biologically relevant relations, as suggested by evidence from the literature.

The model has been applied only to one specific context and thus a proper assessment needs to be made of its application to other datasets based on other contextual backgrounds. It would be useful to apply the methodology with no specific background, potentially capturing global patterns prevalent in the data.

There is also significant improvement that can be made in terms of the model itself. Alternative sampling schemes may better explore the complicated posterior space, and thus potentially uncover richer structure from the dataset. Longer runs of the MCMC sampling scheme may also be useful. This may ensure that the parameter search space is fully explored and that we have not simply sampled from one of many local optima.

5.5 List of References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410.

[2] Aris, J.P., Alvers, A.L., Ferraiuolo, R.A., Fishwick, L.K., Hanvivatpong, A., Hu, D., Kirlew, C., Leonard, M.T., Losin, K.J. and Marraffini, M. (2013). Autophagy and leucine promote chronological longevity and respiration proficiency during calorie restriction in yeast. *Experimental gerontology*.

[3] Aris, J.P., Fishwick, L.K., Marraffini, M.L., Seo, A.Y., Leeuwenburgh, C. and Dunn Jr, W.A. (2012). Amino acid homeostasis and chronological longevity in *Saccharomyces cerevisiae*. In: *Aging Research in Yeast*, pp. 161–186. Springer.

[4] Baldi, P. and Long, A.D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, vol. 17, no. 6, pp. 509–519.

[5] Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K. and Jaakkola, T.S. (2003). Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10146–10151.

[6] Bektas, M. and Rubenstein, D.S. (2011). The role of intracellular protein< i> o</i>-glycosylation in cell adhesion and disease. *Journal of biomedical research*, vol. 25, no. 4, pp. 227–236.

[7] Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008.

[8] Brauer, M.J., Saldanha, A.J., Dolinski, K. and Botstein, D. (2005). Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Molecular biology of the cell*, vol. 16, no. 5, pp. 2503–2517.

[9] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. (2009). Blast+: architecture and applications. *BMC bioinformatics*, vol. 10, no. 1, p. 421.

[10] Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T. and Schroeder, M. (1998). Sgd: Saccharomyces genome database. *Nucleic acids research*, vol. 26, no. 1, pp. 73–79.

[11] Dahl, D. (2006). Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pp. 201–218.

[12] Easlon, E., Tsang, F., Skinner, C., Wang, C. and Lin, S.-J. (2008). The malate–aspartate nadh shuttle components are novel metabolic longevity regulators required for calorie restriction-mediated life span extension in yeast. *Genes & development*, vol. 22, no. 7, pp. 931–944.

[13] Edgar, R., Domrachev, M. and Lash, A.E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, vol. 30, no. 1, pp. 207–210. Data accessible at NCBI GEO database, accession GEO 21187.

[14] Ferguson, T.S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230.

[15] Galdieri, L. and Vancura, A. (2012). Acetyl-coa carboxylase regulates global histone acetylation. *Journal of Biological Chemistry*, vol. 287, no. 28, pp. 23865–23876.

[16] Garcia-Bustos, J.F., Marini, F., Stevenson, I., Frei, C. and Hall, M. (1994). Pik1, an essential phosphatidylinositol 4-kinase associated with the yeast nucleus. *The EMBO journal*, vol. 13, no. 10, p. 2352.

[17] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, vol. 1, no. 3, pp. 515–534.

[18] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian data analysis*. CRC press.

[19] Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Department.

[20] Geyer, C. (2011). Introduction to markov chain monte carlo. *Handbook of Markov Chain Monte Carlo*, pp. 3–48.

[21] Griffiths, T. and Ghahramani, Z. (2005). Infinite latent feature models and the indian buffet process.

[22] Hagberg, A., Schult, D., Swart, P., Conway, D., Séguin-Charbonneau, L., Ellison, C., Edwards, B. and Torrents, J. (2006). Networkx. high productivity software for complex networks. *Webová strá nka https://networkx. lanl. gov/wiki*.

[23] Hastings, W.K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, vol. 57, no. 1, pp. 97–109.

[24] Hedbacker, K. and Carlson, M. (2008). Snf1/ampk pathways in yeast. *Frontiers in bioscience: a journal and virtual library*, vol. 13, p. 2408.

[25] Ishwaran, H. and Zarepour, M. (2000). Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, vol. 87, no. 2, pp. 371–390.

[26] Kaeberlein, M., Burtner, C.R. and Kennedy, B.K. (2007). Recent developments in yeast aging. *PLoS genetics*, vol. 3, no. 5, p. e84.

[27] Kim, P.M. and Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome research*, vol. 13, no. 7, pp. 1706–1718.

[28] Kim, S., Tadesse, M.G. and Vannucci, M. (2006). Variable selection in clustering via dirichlet process mixture models. *Biometrika*, vol. 93, no. 4, pp. 877–893.

[29] Kruskal, J.B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50.

[30] Lu, J.-Y., Lin, Y.-Y., Sheu, J.-C., Wu, J.-T., Lee, F.-J., Chen, Y., Lin, M.-I., Chiang, F.-T., Tai, T.-Y. and Berger, S.L. (2011). Acetylation of yeast ampk controls intrinsic aging independently of caloric restriction. *Cell*, vol. 146, no. 6, pp. 969–979.

[31] McLachlan, G., Do, K.-A. and Ambroise, C. (2005). *Analyzing microarray gene expression data*, vol. 422. Wiley. com.

[32] Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206.

[33] Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, vol. 44, no. 247, pp. 335–341.

[34] Neiswanger, W., Wang, C. and Xing, E. (2013). Asymptotically exact, embarrassingly parallel mcmc. *arXiv preprint arXiv:1311.4780*.

[35] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, vol. 27, no. 1, pp. 29–34.

[36] Ohlssen, D., Sharples, L. and Spiegelhalter, D. (2007). Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons. *Statistics in medicine*, vol. 26, no. 9, pp. 2088–2112.

[37] Orentreich, N., Matias, J.R., DeFelice, A. and Zimmerman, J.A. (1993). Low methionine ingestion by rats extends life span. *The Journal of nutrition*, vol. 123, no. 2, pp. 269–274.

[38] Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, vol. 18, no. 4, pp. 546–554.

[39] Patil, A., Huard, D. and Fonnesbeck, C.J. (2010). Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, vol. 35, no. 4, p. 1.

[40] Piper, M.D., Mair, W. and Partridge, L. (2005). Counting the calories: the role of specific nutrients in extension of life span by food restriction. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 60, no. 5, pp. 549–555.

[41] Roberts, G.O., Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, vol. 7, no. 1, pp. 110–120.

[42] Rocke, D.M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, vol. 8, no. 6, pp. 557–569.

[43] Sanner, M.F. (1999). Python: a programming language for software integration and development. *J Mol Graph Model*, vol. 17, no. 1, pp. 57–61.

[44] Sethuraman, J. (1991). A constructive definition of dirichlet priors. Tech. Rep., DTIC Document.

[45] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, vol. 13, no. 11, pp. 2498–2504.

[46] Shirra, M.K., Patton-Vogt, J., Ulrich, A., Liuta-Tehlivets, O., Kohlwein, S.D., Henry, S.A. and Arndt, K.M. (2001). Inhibition of acetyl coenzyme a carboxylase activity restores expression of the ino1 gene in a snf1mutant strain of *Saccharomyces cerevisiae*. *Molecular and cellular biology*, vol. 21, no. 17, pp. 5710–5722.

[47] Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, vol. 3, no. 1, p. 3.

[48] Smyth, G.K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420. Springer.

[49] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R. and Lander, E.S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550.

[50] Szabo, A., Boucher, K., Carroll, W., Klebanov, L., Tsodikov, A. and Yakovlev, A. (2002). Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Mathematical Biosciences*, vol. 176, no. 1, pp. 71–98.

[51] Teh, Y.W., Görür, D. and Ghahramani, Z. (2007). Stick-breaking construction for the indian buffet process. In: *International Conference on Artificial Intelligence and Statistics*, pp. 556–563.

[52] Walker, S.G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics Simulation and Computation®*, vol. 36, no. 1, pp. 45–54.

[53] Wang, K., Yang, Z., Liu, X., Mao, K., Nair, U. and Klionsky, D.J. (2012). Phosphatidylinositol 4-kinases are required for autophagic membrane trafficking. *Journal of Biological Chemistry*, vol. 287, no. 45, pp. 37964–37972.

[54] Wang, P.-Y., Neretti, N., Whitaker, R., Hosier, S., Chang, C., Lu, D., Rogina, B. and Helfand, S.L. (2009). Long-lived indy and calorie restriction interact to extend life span. *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9262–9267.

[55] Willey, J.M., Sherwood, L. and Woolverton, C.J. (2011). Microbial genomics. In: *Prescott's microbiology*, chap. 16. McGraw-Hill.

[56] Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L. (2001). Validating clustering for gene expression data. *Bioinformatics*, vol. 17, no. 4, pp. 309–318.

[57] Zheng, Q. and Wang, X.-J. (2008). Goeast: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic acids research*, vol. 36, no. suppl 2, pp. W358–W363.

Chapter 6

Conclusion

We applied three different exploratory analyses in an attempt to delineate the interactions between genes. The problem is made complicated not only by the complex dynamical biological system, but also by the broad scope of the problem. Gene interaction was defined as some relationship that occurs between gene products, the regulatory elements of the gene or if some of the genes regulate others. We have referred to their interaction, not in terms of individual isolated units, but as groups, sets or modules. This definition still leaves a broad scope to the problem. Nevertheless, we applied three novel approaches to attempt to hypothesize connections between genes based on the assumptions of the approach and the approach itself.

The first approach was a targetted approach applied to grapevine gene expression data. The target genes modelled were defence-related genes, and the dataset consisted of several microarray experiments. From the expression data we modelled putative sets of co-expression based on our targets and attempted to infer some association between these sets. We uncovered evidence to suggest that some of our defence-related genes have several functions and they may play a role under various conditions in grapevine. The resulting structure from our approach proposed hypotheses that agreed with evidence in the literature. This literature evidence related to specific target genes, or their sets of putatively co-expressed genes, and their activity in our proposed dominant conditions. The sets themselves, when analysed for biologically relevant associations from the literature, agreed to some extent with evidence in the literature. Based on the assumption that putatively co-expressed genes have some common function or set of functions, we proposed a biologically relavent relation between target genes as well as biologically relavent association for some individual target genes. These relationships can then be tested more rigorously.

Our second approach, applied to yeast, attempted an untargetted approach, similarly attempting to connect sets of putatively co-expressed genes based on the assumption that modules of co-expressed genes have some relation with each other over time. This approach used these modules to define a search

space for element-wise statistical dependence estimation. The resulting structure appeared to putatively capture within-pathway affects. Then, generating highly connected submodules using clique detection we uncovered modules that connect putatively related pathways. In particular we detected activity surrounding an important energy pathway in yeast, suggesting that the yeast cells remained metabolically active. We also uncovered topologies that proposed hypotheses relating to the underlying experiment. These findings agreed with evidence in the literature. It may, however, be that the subset of the data we modelled potentially biased the outcome, though the approach did still uncover putatively meaningful relationships in this subset. One potential concern with this model is that the statistical tests are dependent on the partitioned search space we defined, the larger the number of variables to compare the less significant individual comparisons may become. This may limit the application of the model to particular datasets.

Our third approach, applied to the same yeast dataset, was a sub-clustering approach based on non-parametric Bayesian models. Relationships between genes were determined by their co-occurrence through repeated clustering. This relationship was summarized by analysing the topology of the network resulting from these relations. From our initial investigation of the biological relevance of these structures we found putative theoretical support for the inferred relationships from the literature. The community modules appeared to reflect the underlying experimental condition of the dataset. The relationships between the modules therefore suggested possible hypotheses regarding putative inter-related activities or functions. The model used a large number of parameters and thus attempts to navigate a potentially complicated posterior landscape. This may require a more in-depth investigation into the model, as the convergence achieved may be misleading. The model could potentially be multi-modal in the posterior, thus a longer run needs to be tested. The larger the number of parameters and the longer the chain, the more computational resources are required, which may be inhibitory. The current sampling scheme used is also very naive and alternative more efficient schemes exist that should be investigated.

The methods applied represent only a fraction of the possible models that could be applied to this problem. Each model, depending on its assumptions, may potentially capture a different aspect of the problem at hand. The hypotheses generated here are the result of an exploratory analysis and thus require further investigation. Our models have also been applied each to a specific context, applying them to different datasets is an important next step. A deeper mathematical exploration of the models along with their assumptions is also important, as this may uncover new insights or potentially alter existing ones.